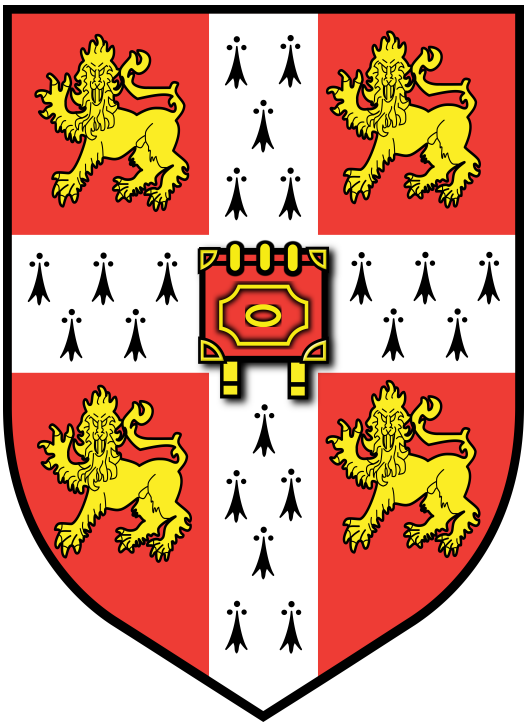
University of Cambridge Saint Catharine's College



Identification and Evolution of New Orthogonal Aminoacyl-tRNA Synthetase/tRNA Pairs for Genetic Code Expansion

This thesis is submitted for the degree of Doctor of Philosophy

by
Daniele Cervettini

March 2020

Preface

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

The research described in this thesis is contained in the following publication:

"Cervettini, D., Tang, S., Fried, S.D. *et al.* Rapid discovery and evolution of orthogonal aminoacyl-tRNA synthetase–tRNA pairs. *Nat Biotechnol* **38,** 989–999 (2020)."

Abstract

Identification and Evolution of New Orthogonal Aminoacyl-tRNA Synthetase/tRNA Pairs for Genetic Code Expansion

by Daniele Cervettini

Genetic code expansion is the branch of molecular biology aiming to expand the repertoire of amino acids which can be incorporated into proteins *in vivo*. A central challenge in expanding the genetic code of cells to incorporate non-canonical amino acids is the scalable discovery of aminoacyl-tRNA synthetase (aaRS)–tRNA pairs (the components of the cellular translational machinery which specify the matching between codons and amino acids) that are **orthogonal** in their aminoacylation specificity. An orthogonal pair is composed of an aaRS which can interact with its partner tRNA, but not with any other tRNAs in the host, and a tRNA which is substrate to its partner aaRS, but not to any other aaRS in the host. In this research, candidate orthogonal tRNAs were identified from millions of sequences by implementing a computational analysis which scored their likelihood to be recognised by the endogenous aaRSs in *E. coli*, our model organism. I then developed a rapid, scalable new *in vitro* approach, named tRNA Extension (**tREX**), to determine the *in vivo* aminoacylation status of tRNAs. Using tREX, 243 candidate tRNAs were tested in *E. coli* and 71 orthogonal tRNAs were identified, covering 16 isoacceptor classes. 23 of those formed functional orthogonal tRNA–cognate

aaRS pairs. By performing additional characterisation and molecular evolution of these newly identified functional pairs, we discovered 5 orthogonal pairs, 3 of which displayed high activity in amber suppression, the technique of choice used to implement genetic code expansion in model organisms. I additionally evolved new amino acid substrate specificities for two pairs. Finally, I use tREX to characterize a matrix of 64 orthogonal synthetase-orthogonal tRNA specificities. This work expanded the number of orthogonal pairs available for genetic code expansion, provided a robust pipeline for the discovery of additional orthogonal pairs, and established a foundation for encoding the cellular synthesis of non-canonical biopolymers.

Chapter I – Introduction

Protein Translation

Proteins represent the most versatile biological polymer and constitute an essential class of molecules which is necessary for any form of life. In spite of the variety of functions they fulfil and of the diversity of organisms in which they are found, all proteins consist of the same 20 building blocks, the set of proteinogenic amino acids. These monomers can be polymerised to form linear molecules thanks to the presence of a carboxylic acid and of an amine moieties in their structure, which can be condensed to form a peptide bond. Organisms store their genetic information, which includes the linear sequence for their proteins, by means of nucleic acid sequences, most commonly DNA, as elucidated thanks to famous experiments by Avery–MacLeod–McCarty[1] first and Hershey-Chase[2] later.

The set of rules which define the correspondence between the genetic information contained into

DNA and the distinctive sequence of proteins is named **genetic code**, and the set of molecular processes involved in the relay of such information have been elucidated by memorable experiments during the second half of the 20th century.

Immediately after the discovery of the DNA structure by Watson&Crick[3], hypotheses were brought forward about the mechanism by which genetic information is interpreted to produce proteins in the cells. In a famous symposium[4] Crick postulated in 1958 that the primary sequence of DNA should be a simple code for the amino acid sequence of a particular protein and that an adaptor molecule should exist, perhaps made of RNA, which should carry the amino acids to the template and mediate the transfer of information from nucleic acids to proteins. Furthermore, he theorised that enzymes could be responsible to join a specific amino acid to a specific adaptor or set of adaptors.

In the following years, all of Crick's suppositions were confirmed to be true. Nirenberg and co-workers first showed that a stretch of poly-uracil can guide the synthesis of a poly-phenylalanine peptide, highlighting the role of RNA in the protein synthesis process[5], then cracked the triplet-based genetic code[6]. The identity of the adaptor molecules was identified by Zamecnik and co-workers, who showed that some soluble RNA (sRNA) molecules were covalently bound to ^{14}C-labelled amino acids in the presence of ATP and amino acid-activating enzymes[7], in a crucial step for protein synthesis[8, 9]. The connection between the genetic information stored in the DNA and the RNA used to synthesise protein was found when the key enzyme, RNA polymerase, was purified by S. Weiss and L. Gladstone from mammalian cell extracts[10, 11].

From the studies listed above, we now know that the DNA information is first copied into RNA in a process known as **transcription**, and that RNA is directly used as a template for protein synthesis. Groups of three letters of RNA, known as **codon**, uniquely correspond to a specific amino acid and the sRNA adaptor molecules which mediate this matching between codons and amino acids are now known as **tRNAs**. These adaptors are recognised specifically by a family of activating enzymes known as **aminoacyl-tRNA synthetases**, or aaRSs, which are responsible to ensure that the correct amino acid is chemically bound to the right tRNA.

In the cell, the molecular machinery responsible to coordinate the conversion of genetic information to functional protein is the ribosome. It consist of a multi-molecular complex composed of RNAs and proteins which moves along mRNAs by one codon at the time, ensuring the correct matching between codons and tRNAs to guarantee high fidelity in the protein synthesis process, known as **translation**12.

In the next sections I will describe in more detail the structure and function of tRNAs and aaRSs, how they interact with each other and how ribosomal protein translation is carried out.

tRNAs

In all organisms tRNAs are key molecules responsible to guarantee the correct flow of information between DNA and proteins. They consist of highly modified and strongly structured RNA molecules[13] whose length varies between 70 to 130 nt. tRNA primary structure is characterised by a high degree of divergence, both within an organism and among different organisms[14]. In spite of this divergence in sequence, tRNAs present a common secondary structure, which is best known as cloverleaf structure. The characteristic features of the vast majority of tRNAs can be described taking *E. coli* tRNAs as an example[15] (**Figure 1.1a**):

i) the two ends of tRNAs come together to form the acceptor stem;

ii) the three nucleotides at the 3'-end are universally conserved and have sequence 5'-CCA-3';

iii) the three nucleotides responsible for the interaction with codons in mRNAs, known as **anticodon**, are present in the middle of a 7 nt loop (anticodon loop) connected to a 5 bp stem (anticodon stem). Together, the anticodon stem and the anticodon loop form the **anticodon arm**;

iv) a structure composed of 3 to 4 bp stem and a loop of variable length is present upstream of the anticodon arm. This structure is called **D arm** due to the presence of dehydrouridine in the loop[16];

v) a **variable loop** of length varying between 3 and 21 nt is present downstream of the anticodon arm and is called variable arm. This structure may or may not fold in a double-stranded stem;

vi) downstream of the variable arm, the last secondary structure element is represented by a 5 nt long stem terminating in a 7 nt loop containing the conserved TΨC motif and is hence called **TΨC arm**, where Ψ stands for pseudouridine.

Using the secondary structure as a reference, *E. coli* tRNAs can be aligned to a standard prototype whose nucleotides are numbered to define the canonical numbering scheme, as indicated below15:

1→7	Acceptor stem, first strand	**39→43**	Anticodon stem, second strand
8→9	Unpaired bases	**44→48**	Variable loop
10→13	D stem, first strand	**49→53**	TΨC stem, first strand
14→21	D loop	**54→60**	TΨC loop
22→25	D stem, second strand	**61→65**	TΨC stem, second strand

26	Unpaired base	**66→72**	Acceptor stem, second strand
27→31	Anticodon stem, first strand	**73**	Discriminator base
32→38	Anticodon loop	**74→76**	CCA end

In *E. coli* some positions are conserved among all tRNAs[15] (**Figure 1.1a**). These include canonical positions 8 which is always U, and positions 74→76, which have sequence CCA. Furthermore, the D loop always contain two Gs which define positions 18 and 19, while the TΨC loop begins with the conserved triplet TTC, which defines positions 54→56 (**Figure 1.1a**).

In addition to possessing well defined secondary structure elements, tRNAs adopt a conserved tertiary structure[18]. In particular, the D arm and the TΨC fold back to establish mutual interactions which confer a characteristic L shape to tRNAs (**Figure 1.1b**). This three-dimensional conformation is approximately retained even in tRNAs where the secondary structure differs from the canonical one described, and guarantees that tRNAs can be recognised correctly by the ribosome[19].

Overall, divergence at the primary structure allows different tRNAs to possess a distinctive fingerprint, which is necessary for specific interactions to occur[20]. As an example, aminoacyl-tRNA synthetases, which are responsible for chemically linking tRNAs with their corresponding amino acid, take advantage of this divergence to engage a specific subset of the cellular tRNA pool[20]. Conversely, the existence of a conserved tertiary structure ensures that distinct tRNAs can establish interactions

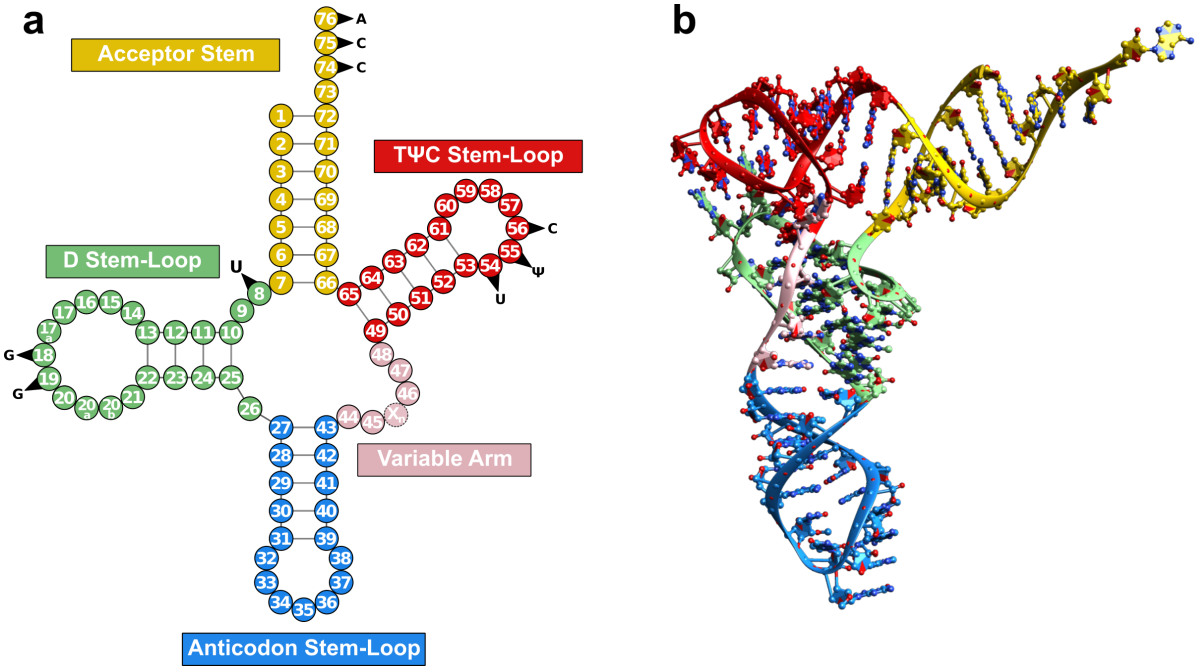


Figure 1.1: a) *Schematic representation of tRNA secondary structure with canonical numbering scheme. Conserved positions among E. coli tRNAs are indicated.* ***b)*** *Tertiary structure of tRNA^{Phe} from S. cerevisiae (PDB 1ehz[17]) adopts the canonical L shapes. Secondary structures are colour coded as in* ***a)****.*

with common partners involved in the translational apparatus.

Aminoacyl-tRNA Synthetases

tRNAs can fulfil their role of adaptors between the codes of nucleic acid and amino acids if, in addition to displaying a solvent-exposed anticodon able to base-pair with a codon in an mRNA, they can be uniquely linked to a specific amino acid. In the cell, a set of 20 enzymes is responsible for catalysing the ATP-dependent esterification of an amino acid's carboxylic moiety onto one of the two hydroxyl groups at the 3'-end of tRNAs. By discriminating a specific subset of tRNAs within the cellular pool, and being able to use only one amino acid as a substrate, these enzymes, the **aminoacyl-tRNA synthetases** (aaRSs), are responsible to establish the correspondence between the alphabet of DNA and the alphabet of proteins, which defines the **genetic code** of an organism. Individually, these enzymes are indicated by the three letter code for their amino acid substrate (e.g.: AlaRS: alanyl-tRNA synthetase).

aaRSs catalyse a reaction which proceeds in two steps: first, the carboxylic acid attached to the α-carbon of amino acids is activated to form an aminoacyl-AMP intermediate. In the second step, the activated carboxylic acid is transferred to either the 2'-OH or 3'-OH of the terminal adenine of the tRNA and the products are released[21]. While protein synthesis requires aminoacylation on the 3'-OH, rapid trans-esterification occurs which causes co-existence of both species for all tRNAs[22].

Based on structural features, aaRSs are commonly divided in two classes[21]. Class I synthetases are defined by the presence of a Rossmann fold, a structural element involved in ATP binding, containing two conserved motifs with sequence "HIGH" and "KMSKS". These aaRSs are mostly monomeric, with few exceptions which form dimers, they approach the acceptor stem of their substrate tRNA from the minor groove and charge its 2'-hydroxyl group. Conversely, class II synthetases lack the Rossmann fold and instead present a 7-stranded β-sheet fold involved in ATP binding, are mostly dimeric or tetrameric and charge the 3'-hydroxyl group (with the exception of phenylalanyl-tRNA synthetase).

aaRSs ensure high fidelity of protein synthesis by preventing charging of the wrong amino acid onto their substrate tRNAs. The low error rate of aminoacylation (of the order of 10^{-4})[23] is maintained by means of two strategies. Firstly, synthetases have evolved to establish specific interactions with only one amino acid[24]. For charged and polar amino acids, for example, a distinctive hydrogen bonds network can only be established between the enzyme and the side chain of one amino acid, ensuring binding of a specific substrate for catalysis. For hydrophobic amino acids, non-polar interactions in

combination with size exclusion of substrates with the incorrect side chain can provide specificity. In certain cases, however, these strategies are not enough to provide a significant difference in binding energy among different substrates. As an example, isoleucyl-tRNA synthetase (IleRS) must prevent mis-charging of $tRNA^{Ile}$ by valine[25]. As both amino acids are small and hydrophobic, no specific hydrogen bond patterns can discriminate between them. Furthermore, IleRS cannot exclude valine by size because valine is one methyl group smaller than isoleucine. In these cases, aaRSs exploit a post-transfer correction mechanism known as **editing**, which hydrolyses incorrect aminoacyl-tRNA complexes[25]. In fact, as isoleucine, being larger than valine, can be actively excluded from the active site of the editing domain of IleRS by size[25], $Val\text{-}tRNA^{Ile}$ can be selectively hydrolysed while preserving the pool of $Ile\text{-}tRNA^{Ile}$.

Aminoacyl-tRNA Synthetase Interactions with tRNAs

In addition to showing selectivity for one amino acid, aaRSs must be able to recognise only a subset of the cellular tRNAs. tRNAs which are recognised by the same aaRS are known as **isoacceptor** tRNAs, as they get charged with the same amino acid. Isoacceptor tRNAs are indicated with the three-letter code of the amino acid which they encode and in addition their anticodon can be indicated (e.g.: $tRNA^{Ala}$ is one of the alanyl isoacceptor tRNAs, $tRNA^{Ala}_{GGC}$ indicates the alanyl isoacceptor tRNA with anticodon GGC). To ensure precise correspondence between anticodons and amino acids, and consequently a precise matching between genetic information and protein sequence, each tRNA must be substrate for one and only one aaRS. In spite of the high levels of conservation of their three-dimensional shape, tRNAs diverge at their primary sequence[15], resulting in heterogeneity in the chemical moieties exposed to the solvent through in the major and minor groove, as well as in the nucleotides not engaged in base pairing. As a result, aaRSs can effectively test the identity of a tRNAs by sampling a discrete number of spots along their interaction surface[20]. If a tRNA displays the correct pattern of chemical moieties at those locations, correct binding occurs and catalysis is favoured. Conversely, tRNAs which don't establish the correct set of interactions with an aaRS do not get effectively aminoacylated.

For each tRNA, the nucleotides that are specifically recognised by its aaRS and which mediate their interaction are known as **identity elements**[20]. Understandably, these positions tend to be conserved among tRNAs belonging to the same isoacceptor class within an organism and among closely related species[26]. A key feature which guarantees fidelity of the genetic code is the difference in the identity elements recognised by different aaRSs. As an example, PheRS is the only one which recognises the top part of the anticodon stem and some nucleotides in the TΨC arm , while LeuRS is among the few

which recognise features of the variable arm[20]. For those positions which are identity elements to multiple aaRS, different enzymes recognise different sequences. As an example, the anticodon represent a key interaction partner for 17 out of 20 aaRSs[20] (AlaRS, LeuRS and SerRS being the exception), however, each of those synthetases is highly specific for the anticodon of its isoacceptor tRNAs, such that variations at those positions are poorly tolerated. Being in close proximity to the catalytic site, the acceptor stem contains identity elements for a large number of aaRSs. Some synthetases also recognise the presence of a variable arm which forms a stem-loop structure (e.g.: SerRS) or other unusual features, like the presence of one extra nucleotide at the 5'-end of the tRNA which base pairs with canonical position 73, which is usually unpaired (e.g: HisRS)[20].

The existence of a variety of interaction points distributed along the whole tRNA body is responsible for the affinity of each aaRS for tRNAs belonging to its isoacceptor class and not others. In addition, the coexistence of all aaRSs in the cell generate a competition for their respective substrates. Taken together, these mechanisms guarantee fidelity of tRNAs aminoacylation, a key requirement for accurate protein synthesis.

Ribosomal protein synthesis

Aminoacylated tRNAs are the building blocks for translation. In the cell, a complex apparatus exists which ensures correct processing of this raw material into the correct final product. A key character of such apparatus is the **ribosome**, a megadalton-size ribonucleoprotein which monitors the correct matching between mRNA codons and tRNA anticodons and catalyses the condensation reaction between amino acids to form polypeptides[12].

The prokaryotic ribosome is constituted of two subunits of different sizes. The small subunit, or 30S, is composed of a single ribosomal RNA (16S) associated to 21 proteins. The large subunit, or 50S, consists of two different RNAs, the 5S and the 23S rRNAs, and 31 additional proteins. When associated together, the two subunits for the 70S ribosome. Both subunits present three distinct sites which can accommodate tRNAs at different stages of the translation process, known as A site, P site and E site.

Translation initiation starts when the 3'-end of the the 16S rRNA of the small subunit, which has sequence 5'-ACCUCCUUA-3' in *E. coli*, forms a base-pair interaction with a sequence of an mRNA known as **Shine-Dalgarno** (SD) sequence (with canonical sequence AGGAGG). When this interaction is established, the translation start codon ATG is displayed in the P site of the small subunit. A specialised initiator tRNA, fMet-tRNA^{fMet}, charged with an N-formylated methionine (fMet), is escorted to the P site bound to the initiation factor IF2, a protein with GTPase activity. In addition,

initiation factors 1 (IF1) and 3 (IF3) are occupy the A and P sites, respectively, in order to stabilise the initiation complex as well as preventing premature association of the large subunit or of incoming tRNAs[12].

GTP hydrolysis by IF2 leads to release of the initiation factors, association of the large subunit and transition to the elongation phase. At the beginning of this phase, the P site is occupied by the fMet-tRNA^{fMet} and the E and A sites are empty. In order for protein synthesis to proceed, a new tRNA is delivered to the A site of the ribosome bound to a small GTPase known as **Elongation Factor, Thermally unstable** (EF-Tu). If the tRNA anticodon correctly base pairs with the codon displayed in the A site, EF-Tu hydrolyses GTP and the tRNA is free to reposition its 3'-end into the peptidyl-transferase centre (**PTC**), the site of the ribosome involved in peptide bond formation. At this stage, the P-site tRNA is deacylated and the peptide chain is transferred to the A-site tRNA, leaving a free tRNA in the P site and a peptide-bound tRNA in the A site. To allow the cycle to be repeated, EF-G triggers translocation of the free P-site tRNA to the E site, where it is eventually ejected, and translocation of the peptidyl-tRNA to the P site. Furthermore, the ribosome is moved along the mRNA by one codon[12].

The ribosome possesses proofreading mechanisms which ensure that elongation can only take place if the correct tRNA is present in the A site. In order to do this, some universally conserved position in the 16S rRNA (G530, A1492 and A1493 in *E. coli*) form a hydrogen bond network with the first two positions of the codon-anticodon base pairs which only allows Watson-Crick interactions in order for elongation to proceed. At the third position of the codon-anticodon base pairs, a looser level of control allows wobble interactions to occur, resulting in a tRNA to be able to decode more than one codon[12].

The elongation process repeats until a stop codon is found in the A site. Physiologically, no tRNAs are present in the cell whose anticodon can base pair with the three codons UAA, UAG or UGA. Instead, a release factor binds to the A site and leads to hydrolysis of the polypeptide from the P-site tRNA, hence termination of translation. In bacteria, release factor 1 (RF1) and 2 (RF2) recognise stop codons UAG and UGA, respectively, while both can recognise the UAA codon. Following termination, the ribosome is disassembled by the action of a variety of GTPases, including RF3, RRF and EF-G and recycled for new rounds of translation[12].

Genetic Code Expansion

In each living organisms all 64 possible RNA codons are assigned to either an amino acid (sense codons), by the presence of a tRNA with a matching anticodon and an aminoacyl-tRNA synthetase that can charge it, or to termination of translation (stop codon), by a corresponding release factor. The set of rules which determine the meaning of codons for protein translation, or **genetic code**, is almost universally conserved, even if few minor differences have been identified in various organisms[27]. The branch of molecular biology whose aim is to increase the variety of amino acid which can be incorporated into proteins by means of the cellular translational machinery is called **genetic code expansion**[28]. In order to achieve this result in a living cell, a number of conditions must be satisfied:

i) a codon must be assigned to the newly introduced amino acid;

ii) a tRNA capable of decoding the chosen codon must be available;

iii) an aminoacyl-tRNA synthetase must exist which can selectively charge this tRNA with the amino acid of choice.

Importantly, newly introduced amino acids, referred to as **non-canonical amino acids** (ncAAs) have been shown to take part in the ribosomal peptide bond synthesis, even if with an efficiency dependent on their structure[29]. Consequently, if the above conditions are met, the cell will start interpreting the occurrences of the re-specified codon as the ncAA.

However, to make genetic code expansion possible, some additional considerations must be taken into account. Fidelity of protein translation is ensured by the lack of cross-reactivity between cellular aaRSs and non-cognate tRNAs from other isoacceptor classes. When introducing new aaRS/tRNA pairs in a cell, in order to prevent repercussions on cellular viability, this fidelity should not be compromised. In fact, erroneous mis-incorporation of a ncAA across the cellular proteome would result in the production of aberrant proteins with impaired function and in the reduction of cellular viability[30]. To prevent this erroneous mis-incorporation of a ncAA, the aaRSs able to use it as a substrate, which we may call **ncAA-RS**, must not use cellular tRNAs as substrates, generating a pool of mis-charged tRNAs. Enzymes with this property are referred to as **orthogonal**, meaning they cannot recognise and aminoacylate the tRNA pool of the organisms in which genetic code expansion is performed, while having their cognate tRNA ($tRNA^{ncAA}$) as a sole substrate. Conversely, a $tRNA^{ncAA}$ is orthogonal if it cannot be a substrate for any cellular aaRS, while being recognised by its ncAA-RS.

tRNA orthogonality is required to ensure accuracy in the incorporation of the ncAA at the target codon.

The possibility to expand the endogenous pool of proteinogenic amino acid is attractive on many regards[28, 31]. In fact, it should be noted that, in spite of the enormous variety of functions accomplished by proteins, they are built from an extremely limited selection of chemical moieties. Site-specific introduction of new side chains can, for example, expand the variety of reactions which proteins can undergo or catalyse. As an example, the introduction of groups which can take part in bio-orthogonal reactions opens up new ways of labelling proteins to study their physiology, or of purifying them with new methods[32, 33]. Amino acids containing photo-reactive groups which can be activated in response to light stimuli can be exploited to induce perturbations with high spatiotemporal control and resolution. Such ncAAs have been used to study protein interactions[34-36], to alter cellular physiology[37, 38] etc.

Another area in which genetic code expansion can be a powerful tool is the study of post-translation modifications (PTMs). In fact, introducing site-specific PTMs is relatively difficult, as the enzymes involved often have a broad specificity[39] or are unknown. On the other hand, the introduction of ncAA containing PTMs can be performed regardless of any knowledge on the enzymes involved in the physiology of that PTM and without interference on other proteins, providing superior resolution. To date, several PTMs, like serine phosphorylation[40, 41], threonine phosphorylation[42], lysine acetylation[43], lysine methylation[44] etc.[28, 31], have been introduced by GCE.

Incorporation of a ncAA only at the desired site of interest is difficult to achieve *in vivo*. Even when an orthogonal ncAA-RS is available which is completely specific for its $tRNA^{ncAA}$, the latter might decode a codon which occurs in locations other than the site of interest of the target protein. This is especially problematic considering that no codons are naturally unassigned to be either sense or stop codons, with no blanks available. In the next sections early attempts to expand the genetic code will be discussed to highlight how the lack of available blank codons can be overcome and how orthogonal aaRS/tRNA pairs can be obtained.

Amber Suppression

The TAG codon was the first one identified to induce termination of protein synthesis in mutants T4 phages irradiated with UV light. Some mutants were identified which expressed a truncated form of the protein containing the mutation, called nonsense mutation[45]. Furthermore, truncation was observed when the mutated protein was expressed in some strains of *E. coli* while other strains were

able to suppress such mutation. The "suppressable" mutations were termed *amber*[46]. The molecular cause of the truncation was the appearance of the TAG codon along the protein coding sequence, which for this reason is known as amber stop codon. Furthermore, the genes responsible to suppress the amber mutation in the several suppressor strains of *E. coli* were pinned down as mutant tRNAs whose anticodon had mutated to 5'-CUA-3' [47]. As a result, these tRNAs could decode the TAG codon, preventing premature termination and restoring production of the full length protein.

The possibility of suppressing a stop codon with an appropriate tRNA implies that the activity of the release factor 1 can be counteracted to some extent. In addition, among all the stop signals, the amber codon is the one used least frequently in many organisms, including *E. coli*, *S. cerevisiae, C. elegans* and also in humans[48]. Given its good efficiency, amber suppression is the technique which allows the site-specific incorporation of a ncAA of interest with the highest precision, provided an appropriate ncAA-RS/$tRNA^{ncAA}$ is found.

Early Work on Genetic Code Expansion

In some early attempts to direct site-specific incorporation of ncAA into proteins, Furter transplanted the PheRS/$tRNA^{Phe}$ pair from *S. cerevisiae* (*Sc*), which is naturally able to use *p*-fluoro-phenylalanine as a substrate, into an *E. coli* strain incapable of using the ncAA as a substrate for the endogenous synthetases[49]. In his experiments, the anticodon of the *Sc*-$tRNA^{Phe}$ was mutated to CUA (*Sc*- $tRNA^{Phe}_{CUA}$). By using a reporter gene containing an in-frame amber stop codon, Furter observed incorporation of *p*-fluoro-phenylalanine with ~70% yield, with the remaining fraction being phenylalanine and lysine. His experiment highlighted that while mutations of the tRNA anticodon from its wild type sequence to CUA did not disrupt the recognition by *Sc*-PheRS, it resulted in a $tRNA^{Phe}_{CUA}$ which was not orthogonal in *E. coli*, being charged by the *E. coli* LysRS.

To improve incorporation accuracy, later works by the Schultz lab focused on the identification of fully orthogonal pairs for genetic code expansion. Once identified, orthogonal pairs could be evolved to alter their recognition for a substrate amino acid and redirect it towards a ncAA. In their first attempt, they tried to generating an orthogonal pair by engineering the *E. coli* GlnRS/ $tRNA^{Gln}_2$ [50]. While evolution of the $tRNA_2^{Gln}$ was successful in generating a tRNA which was orthogonal to the wild type GlnRS, no variants of the synthetase could be found which did not retain any specificity towards the wild type $tRNA^{Gln}_2$.

A more fruitful strategy to generate orthogonal pairs was found to be the transplantation of

aaRS/tRNA pairs found in other organisms which have evolved distinctive features compared to the orthologue pairs in the host, conferring them orthogonality in the recipient host. The first of these examples was the TyrRS/$tRNA^{Tyr}$ derived from *Methanocaldococcus jannaschii* (*Mj*), whose genome was fully sequenced at the end of the 1990s[51]. Unlike its *E. coli* counterpart, this archaeal tRNA does not display an extended variable arm and has a C1:G72 base pair, which differs from the bacterial identity element G1:C72. In addition, the *Mj*− $tRNA^{Tyr}_{CUA}$ is orthogonal in *E. coli* and is charged by its cognate *Mj*-TyrRS with tyrosine[52].

The Schultz lab envisioned an experimental strategy to alter the amino acid specificity of the *Mj*-TyrRS by directed evolution[53]. First, a large library of mutant synthetases (>10^8 mutants) was built by randomizing the residues of the enzyme lining the amino acid binding pocket. A selection marker was generated by introducing an amber stop codon in a permissive site of the chloramphenicol acetyl-transferase gene (cat^{112*}). If cells expressing cat^{112*} and *Mj*− $tRNA^{Tyr}_{CUA}$, are transformed with the library and grown on chloramphenicol- and ncAA-containing media, only colonies harbouring an active synthetase variant could survive. This positive selection strategy is not capable of discriminating synthetases which incorporate the ncAA from others which recognise other cellular amino acids. To circumvent this limitation, a negative selection was performed using a toxic gene interrupted by an amber stop codon instead of cat^{112*}. When the selection was performed using the same setup but using the toxic marker in place of the antibiotic resistance marker and in the absence of ncAA in the medium, synthetase variants specific for other cellular amino acids were filtered out. By alternating multiple consecutive rounds of positive and negative selection, the Schultz group initially identified a synthetase specific for *O*-methyl-tyrosine, while tens of other enzymes were evolved by a variety of groups using comparable strategies to incorporate a multitude of ncAAs. To date, this *Mj*-TyrRS/$tRNA^{Tyr}$ system remains one of the most successful one employed for genetic code expansion[54]. However, its active site is mostly limited to accommodating tyrosine derivatives[54].

The Pyrrolysyl-tRNA Synthetase/tRNA Pair

A major breakthrough for the advancement of genetic code expansion came when studies about methanogenesis in some bacteria belonging to the *Methanosarcina* genus identified an in-frame amber stop codon within the *mtmb1* gene coding for monomethylamine methyltransferase[56, 57]. The X-ray diffraction pattern of crystals grown from the purified protein identified an electron density corresponding to a new amino acid in the active site of the enzyme[58]. Subsequent studies elucidated the chemical structure and biosynthetic pathway which produced the new amino acid, named **pyrrolysine** (Pyl), which is fundamental for methanogenesis due to its unique chemical properties[59].

These studies also clarified that pyrrolysine was not synthesised as a post-translational modification, but instead it was incorporated within the protein sequence co-translationally thanks to the action of a unique tRNA, $tRNA^{Pyl}$, encoded by the *pylT* gene, and a class II aminoacyl-tRNA synthetase, PylRS, encoded by the *pylS* gene, located in the same operon as the enzyme involved in pyrrolysine biosynthesis and composed of an N-terminal domain responsible for tRNA binding and a catalytic C-terminal domain which performs aminoacylation[60]. Importantly, $tRNA^{Pyl}$ naturally possesses the 5'-CUA-3' anticodon required to perform amber suppression, and its interaction with the translational machinery is not mediated by any additional components compared to the canonical tRNAs[60]. In this regard, pyrrolysine can be considered a case of natural expansion of the genetic code.

The PylRS/$tRNA^{Pyl}$ system presents some remarkable features. Unlike any of *E. coli* tRNAs, $tRNA^{Pyl}$ from *Methanosarcina barkeri* (*Mm*) or *mazei* (*Mb*) presents a single nucleotide in between the acceptor stem and the D stem, a D loop composed of only 5 nucleotides and lacking one of the conserved GG at canonical positions 18-19, an acceptor stem composed of 6 base pairs instead of 5 and a variable loop composed of only 3 nucleotides (**Figure 1.2a**). These characteristics render the *Mm*/*Mb*-$tRNA^{Pyl}$ orthogonal in *E. coli* and indeed in other model organisms[28]. Furthermore, as *Mm*/*Mb*-PylRS has adapted to recognise such an uncommon tRNA (**Figure 1.2b**), it displays no reactivity towards endogenous tRNAs of those same species. These two evidences define the PylRS/$tRNA^{Pyl}$ pair as orthogonal. Furthermore, like only few other aaRSs, PylRS does not interact with the anticodon of its

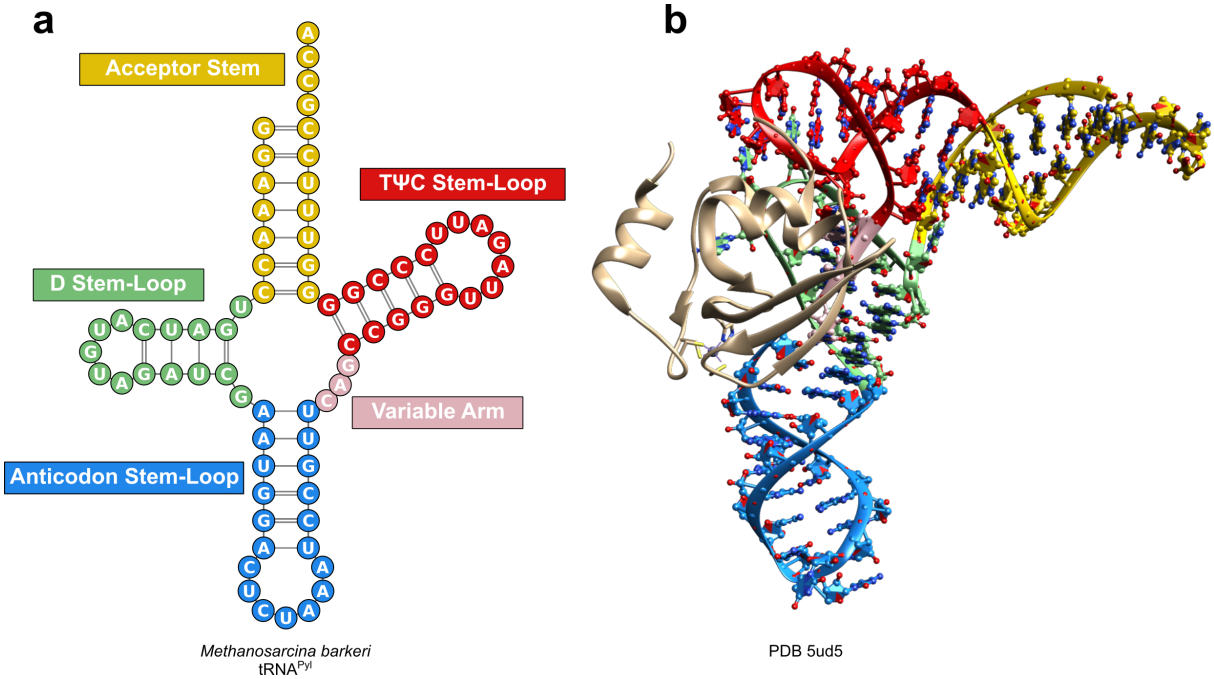


***Figure 1.2**: **a**) Secondary structure of the tRNAPyl from M. barkeri highlights the unusual characteristics of this tRNA, such as the lack of canonical position 9, the smaller size of the D loop, the extension of the anticodon arm from 5 to 6 bp and the reduction of the variable loop to only 3 nt. **b**) Crystal structure of the N-terminal domain of the PylRS from M. mazei (PDB 5ud5[55]) highlights its close proximity to the variable loop of the enzyme, which is smaller than the variable loop of any of the E. coli tRNAs.*

cognate $tRNA^{Pyl}$. Instead, the N-terminal domain of the enzyme makes contact with the variable arm, while the catalytic domain contacts the acceptor stem and the D arm[55, 61].

The active site in PylRS is naturally promiscuous and can tolerate other lysine derivative beyond pyrrolysine[57, 62]. Moreover, site-directed mutagenesis of the active site allowed incorporation of a very broad range of ncAA using this system[57]. Among these, a large number of lysine derivatives or analogues are present, but some mutants are capable of recognise aromatic ncAA deriving from phenylalanine[57].

More recently, the expanding availability of genomic data from bacterial and archaeal species allowed the identification of a new class of PylRS/$tRNA^{Pyl}$ pairs. Interestingly, this class in characterised by synthetases lacking the N-terminal domain responsible for tRNA binding. Unlike some bacterial PylRSs, this domain is not expressed as an independent protein from a distinct gene, but is completely absent instead[63]. Heterologous expression of PylRS/$tRNA^{Pyl}$ belonging to this class, denoted as ΔN class to reflect the truncation of the N-terminal domain, into *E. coli* confirmed that these synthetases are competent in performing aminoacylation and that the tRNAs are active amber suppressors. While displaying natural cross reactivity among themselves, the heterogeneity among members within the ΔN class and between the two different classes allowed engineering of mutually orthogonal pairs64.

The Phosphoseryl-tRNA Synthetase/tRNA Pair

In spite of its rather limited substrate scope, an important pair used for genetic code expansion is represented by the phosphoseryl-tRNA synthetase (SepRS) in combination with its cognate $tRNA^{Sep}$. Being one of the most common and pleiotropic post-translation modification, gaining insight into the effect of phosphorylation events at different serines of various targets is an important research topic. However, research into this PTM is limited by the polyspecificity of protein kinases[39] and by limited understanding of the enzymes responsible for phosphorylation at specific sites of target proteins.

An important breakthrough in the production of phosphorylated proteins by genetic code expasion came from the observation that several archaeal genomes, such as the ones from *Methanocaldococcus jannaschii*, *Methanobacterium thermoautotrophicum* and *Methanococcus maripaludis*, lacked any gene which could be associated to a cysteinyl-tRNA synthetase, in spite of its existence in all other organisms known to date and in spite of the presence of cysteine in the proteome of those archaea[65]. Several hypotheses were formulated to explain the biogenesis of cysteine incorporation in those species lacking CysRS, including the involvement of ProRS in the process or the existence of an

unusual new type of CysRS in those archaea[65]. However, in 2005 the Söll group proved that synthesis of Cys-tRNA^{Cys} in *M. jannaschii* is a two-step process which starts with the acylation of tRNA^{Cys} by a new homotetrameric class II aaRS capable of using *O*-phosphoserine (Sep) as a substrate, to form Sep-tRNA^{Cys} 66. Subsequently, a PLP-dependent enzyme is able to convert Sep-tRNA^{Cys} into Cys-tRNA^{Cys} by using a sulphur donor to replace the phosphate group of Sep.

Following identification of this novel mechanism for cysteine synthesis and incorporation in some archaea, the biochemistry of the interaction between SepRS and Sep or tRNA^{Cys} was investigated. The crystal structure of the enzyme revealed that phosphoserine recognition occurs by means of a precise hydrogen bond network established between the phosphate group and the enzyme, without the involvement of charge-to-charge interactions[67]. The formation of a large number of weak interaction, compared to stronger salt bridge interactions, ensures high fidelity for the correct substrate over isosteric analogues, like glutamic acid, which is present in the cell at a very high concentration (e.g. up to 100 mM in *E. coli* grown in minimal medium[68]).

The discovery of this natural SepRS/tRNA^{Cys} pair capable of generating Sep-tRNA^{Cys} provided an idea starting point for the expansion of the genetic code with phosphoserine. In 2011, Park *et al.*[40] described the generation of the amber suppressor tRNA^{Sep} by mutating the *M. jannaschii* (*Mj*) tRNA^{Cys} at the anticodon (G34C, C35U) and at position 20 (C20U). This tRNA^{Sep} was not detectably aminoacylated by *E. coli* aaRSs. Co-expression of this tRNA together with SepRS from the mesophilic *Methanococcus maripaludis* in a strain engineered to have higher intracellular concentration of phosphoserine, a natural metabolite for *E. coli*, resulted in the synthesis of Sep-tRNA^{Sep}, but in little or no amber suppression, due to poor delivery of the aminoacylated tRNA to the ribosome. To account for this, the group evolved a mutant of EF-Tu, called EF-Sep, with improved binding to Sep-tRNA^{Sep}, leading for the first time to amber suppression with phosphoserine in *E. coli*.

The system described by Park *et al*. suffered from poor overall amber suppression efficiency, due to the severe loss of activity of SepRS on the amber suppressor mutant of *Mj*-tRNA^{Cys}. Subsequent work substantially improved the overall activity of the pair by engineering the interaction between the enzyme and its cognate tRNA's anticodon[41]. This result was achieved first by mutating the tRNA sequence surrounding the anticodon to identify variants which performed better in amber suppression assays, then by evolving SepRS in its anticodon-binding domain. In the same work, the authors proved that non-hydrolysable analogues of phosphoserine could be incorporated by their improved system as well. Subsequently, orthogonality of tRNA^{Sep} was improved and the active site of SepRS was expanded to accommodate *O*-phosphothreonine[42]. To date, this pair has been used to elucidate new roles for protein phosphorylation in *in vitro* and *in vivo* studies, including in mammalian cells[69].

Incorporation of Multiple ncAA into Proteins

Amber suppression represents a valuable tool for the co-translational incorporation of a single ncAA into a protein. Multiple incorporations of a single ncAA can also be achieved using this same technology if multiple occurrences of the amber codon appear within the same open reading frame. However, due to competition between amber suppression and protein termination by RF1, the production of full-length protein decreases progressively when increasing the number of stop codons[70].

When the experimental setup requires the incorporation of multiple distinct ncAA within the same protein, new limitations need to be taken into account. Firstly, codons availability becomes a fundamental limiting factor. While intuitively suppression of other stop codons might seem reasonable, the higher frequency of TGA and TAA codons in the genome compared to TAG reduces the fidelity of genetic code expansion using those codons, while at the same time reducing the overall efficiency of suppression[71]. In fact, TGA suppression by a mutant $tRNA^{Pyl}$ carrying the UCA anticodon, which does not interfere with aminocylation by PylRS, is overall significantly less efficient in *E. coli* while also suffering from a measurable level of natural suppression by the $tRNA^{Trp}$. Instead, suppression of the TAA codon does not suffer from background suppression from endogenous tRNAs but is the least efficient due to the double competition between the suppressor tRNA and both cellular release factors[71].

As an alternative to suppression of stop codons, new black codons must be generated to allow incorporation of new ncAAs site-specifically. To achieve this, three approaches have been explored which will be describe in the next paragraphs:

i) exploiting quadruplet codons;

ii) generating codons containing non canonical bases;

iii) re-assigning the sense codons at the genome-scale.

Availability of free codons represents a necessary but not sufficient requirement for the incorporation of multiple ncAAs. In addition, a set of aaRSs/tRNA pairs must be available which can use the ncAAs as a substrate and which are both orthogonal to the endogenous aaRSs/tRNA pairs found in the host organism and also to each other. The problem of available orthogonal pairs will be discussed later.

Quadruplet Codons and Quadruplet-Reading Ribosomes

Given the four nucleotides which constitute RNA, a code composed of a combination of three of those nucleotides will give rise to 64 possible words. However, if the code contained 4 nucleotides instead, a large number of new words would be generated which could be used to convey new information. The possibility of using quadruplet codons instead of the universal triplet codons finds its basis on the experimental observation of frameshift suppressor tRNAs containing an extended 8 nucleotide anticodon loop capable of suppressing single nucleotide insertion in *Salmonella* and yeast[72].

Genetic code expansion using quadruplet tRNAs was explored by the Sisido group starting from 1999 in an *in vitro* system[73, 74]. They chemically acylated variants of the yeast $tRNA^{Phe}$ containing an extended anticodon, including AGGU, CGGU, GGGU, CUAU, CCCU, CUCU etc. with ncAAs and confirmed that they were effectively used during protein synthesis. These works also highlighted the lack of cross-talk between different quadruplet codons, which was relevant to incorporate two distinct fluorescent amino acids at two sites of calmoduline for FRET measurements[75].

Later, examples of *in vivo* applications of genetic code expansion using quadruplets were shown by Anderson *et al.*[76] in 2004. Building on the knowledge that frameshift suppressors could be derived for $tRNA^{Gln}$ and $tRNA^{Lys}$ in *E. coli* by expansion of the anticodon loop, they generated a quadruplet $tRNA^{Lys}$ with anticodon UCCU derived from *Pyrococcus horikoshii* and evolved it cognate synthetase to incorporate homoglutamine. By combining this new pair with the mutant of *Mj*-TyrRS incorporating *O*-methyl-tyrosine in response to the amber codon, they observed double incorporation of ncAAs into myoglobin with a yield of ~15% compared to the wild type protein variant.

In spite of the possibility to perform quadruplet decoding, the wild type ribosome is not proficient at it. Given its fundamental role in sustaining life, however, mutations affecting its activity are generally poorly tolerated in living organisms. In order to circumvent this limitation, Rackham *et al.*[77] envisioned a strategy to generate an independent copy of the ribosome not involved in the translation of cellular messengers, but directed instead towards a distinct subset of mRNAs, thus generating an orthogonal translation system. To achieve this result, they hypothesised that mutations might exist of the Shine-Dalgarno sequence of an mRNA which could prevent the binding of the small subunit of the cellular ribosomes to it. At the same time, they hypothesised that an orthogonal small subunit might contain mutations at the 3'-end of the 16S rRNA which rescues its binding to the mutant Shine-Dalgarno sequence while minimising interactions with the wild type sequence. To achieve this goal, they generated a dual positive-negative selection marker by fusing in-frame the genes for chloramphenicol acetyltransferase (*cat*), an antibiotic resistance marker, and uracil

phosphoribosyltransferase (*upp*), which confers sensitivity to 5-fluorouracil. They generated a library of the SD sequence between positions -13 and -7 upstream of the start codon of the dual selection marker, transformed it in cells and later grew them on 5-fluorouracil. Variants of the SD sequence recognised by the wild type small subunit resulted in the translation of the *upp* gene and cells containing those variants were not viable. Subsequently, the dysfunctional variants were transformed in cells together with a library of mutants of the 3'-end of the 16S rRNA where the nucleotides interacting with the SD sequence were randomised. Cells containing the two libraries were grown on chloramphenicol to select combinations of SD sequence/16S rRNA which led to effective translation. The selection was successful to identify more than one orthogonal SD/anti-SD sequence, an important achievement towards the generation of an orthogonal cellular translation system.

This orthogonal small ribosomal subunit was later used as a starting point to overcome the above mentioned limitations which are intrinsic to stop codon suppression and quadruplet decoding. First, to improve the efficiency of amber suppression and limit the competition by the release factor, Wang *et al.*[78] designed a library of the 16S rRNA in the loop composed of positions 529 to 535, which was observed to be located proximal to the A site where either the tRNAs or RF1 bind. This library was interrogated to test the ability of the variants to outperform the wild type ribosome in amber suppression. To do this, an orthogonal mRNA containing the *cat* gene was constructed where the gene was interrupted by the UAGA quadruplet codon to be decoded by a $tRNA^{Ser2}$ frameshift suppressor with anticodon UCUA, as a proxy for amber codon. A positive selection, using chloramphenicol in a similar fashion compared to what described above, identified Ribo-X, containing only two mutations U531G and U534A, which significantly increases efficiency of read-through of single or multiple amber stop codons to allow efficient site-specific incorporation of ncAAs.

As a last step on engineering for the 16S rRNA, Ribo-X was further optimised for quadruplet decoding using different quadruplets[79]. Neumann *et al.* built several libraries of Ribo-X mutants and performed a similar selection as before, using a reporter *cat* gene interrupted by an AAGA frameshift insertion which could be suppressed by the quadruplet $tRNA^{Ser2}$ with extended anticodon UCUU. They hence identified a variant, named Ribo-Q1, containing two mutations, A1196G and A1197G, which was capable of decoding quadruplet as efficiently as triplets, thus theoretically opening up a very large number of black quadruplet codons for genetic code expansion. As a proof of principle, Neumann *et al.* showed that it was possible to incorporate two distinct ncAAs, *p*-azido-L-phenylalanine and N^{ε}-[(2-propynyloxy) carbonyl]-L-lysine, at two distinct sites of calmodulin, which could be used to perform intramolecular cycloaddition using Cu(I) catalysis.

This set of experiments demonstrated that the limitation in terms of codons availability for genetic code expansion can be effectively circumvented by the use of quadruplets in combination with an

engineered ribosome. Importantly, this result indicates that codons are not the limiting factor for incorporation of multiple ncAAs within one polypeptide. Furthermore, more alternative approaches have been investigated for the same purpose, as discussed below.

Non-Canonical DNA Bases

In spite of the incredible variety of forms in which life on Earth is manifested, all biological entities, from viruses to bacteria, archaea and eukaryotes, share a communal alphabet used to store genetic information, composed of four different nucleotides which can interact by means of hydrogen bonds to form two distinct base pairs (A:T and C:G). This peculiarity has led scientists to investigate whether other types of genetic letters and genetic pairs could be useful candidates to fulfil the processes of replication, transcription and translation which genetic material is required to undergo, or whether only the existing ones were compatible with life as we know it.

Several investigations have been carried out in this direction. As an example, the Brenner lab tried to develop a new pair which would still rely on hydrogen bonds to ensure specificity without introducing erroneous matching. In their first approach[80], they noticed that the pyrimidine-purine scaffold could theoretically accommodate 6 hydrogen bond patterns of donor/acceptor combinations (some of which require repositioning of the nitrogens in the rings), but that only the two patterns "donor-acceptor-acceptor" (C:G) "acceptor-donor-acceptor" (T:A) were used. To add the pattern "acceptor-acceptor-donor" they synthesised the bases isoC and isoG, and showed incorporation into DNA and RNA. In addition, the lab proved that the new isoC-isoG pair could be effectively used in ribosomal translation to guide the incorporation of the non-canonical amino acid iodo-tyrosine[81]. In their experimental setup, the authors generated an mRNA containing a "(isoC)AG" codon and a synthetic tRNA with anticodon "CU(isoG)". When the synthetic tRNA was not aminoacylated, skipping of the isoC base led the generation of a distinct product. However, when the synthetic tRNA was aminoacylated with iodo-tyrosine, incorporation of this amino acid in the nascent polypeptide was more efficient than what could be synthesised in an equivalent experiment involving amber suppression instead.

While being the first non-canonical bases being investigated, isoC and isoG were subject to tautomerisation and hydrolysis, leading to undesired mis-matches. In another design[82], they took into account the evidence previously reported that natural base pairs present an electron density in the minor groove from either the N3 of the purines or from the exocyclic oxygen of the pyrimidines, to create two new letters, simply called **Z** and **P**, which would retain this feature but displaying a hydrogen bond pattern "donor-donor-acceptor", which is distinct from the two natural ones. In their

work, they showed that DNA sequences containing the 6 letters ATGCZP can be replicated by a DNA polymerase *in vitro* and that did not lead to unidirectional loss of the new bases, while still allowing for occasional mutations, which are the substrate for evolution.

A different approach has been taken by the Hirao lab, that tried to identify a new pair which would not rely on hydrogen bonds. To achieve this, they envisioned the use of hydrophobic interaction to replace the natural hydrophilic interactions which characterise base pairing. They first designed the two bases 7-(2-thienyl)-imidazo[4,5-b]pyridine (**Ds**) and pyrrole-2-carbaldehyde (**Pa**)[83] which are respectively bulkier than a purine and smaller than a pyrimidine, and showed that their unusual hydrophobic interaction combined with unfavourable steric interaction with the natural bases allowed them to form a new independent pair, while being effectively used for both replication and transcription *in vitro*. This pair, however, suffered from Ds:Ds mis-matches. In a subsequent work[84], the group generated a new partner for Ds, 2-nitro-4-propynylpyrrole (**Px**), designed to improve shape complementarity and electron potential complementarity, and showed that this new pair could be effectively used for PCR.

Important progress in the attempt to incorporate non natural base pairs in a biological system came from the Romesberg lab. Continuing in the research of pairs which would not take part in hydrogen bonds, the group first investigated the use of nucleotides formed of two fused aromatic rings, called **d5SICS** and **dNaM**[85] and found out that this pair is capable of being used effectively for replication with no bias for the sequence context surrounding the new pair, condition which is required to allow encoding of the genetic material without restrictions. Later investigations from the same lab on an updated synthetic pair of nucleotides, **dNaM** and **dTPT3**, made significant progress on the generation of an organism with an expanded genetic material[86]. To compensate for the lack of biosynthetic pathways for the production of those nucleotides, the group exploited the transporter for nucleotides triphosphate from *Phaeodactylum tricornutum* to overcome the natural lack of import systems for NTPs and dNPTs in *E. coli,* which is required to allow replication and transcription of the DNA containing the non-canonical bases. By doing so, it was possible to demonstrate transcription of a reporter gene containing a synthetic base within the middle position of a serine codon and translation of the same reporter mediated by a $tRNA^{Ser}$ containing the anticodon with the complementary synthetic base. Using a similar setup, incorporation of ncAAs was observed in response to the same non-canonical codon mediated by mutant synthetases and adapted tRNAs derived from the *M. jannaschii* TyrRS/$tRNA^{Tyr}$ pair or from the PylRS/$tRNA^{Pyl}$ pair.

Overall, the progress in the development of additional DNA letters has opened the possibility of generating new free words for the genetic code. While this option is still in an early stage of investigation, expansion of the base alphabet from 4 to 6 would theoretically enable a total of 6^3=216 codons, with a net increase of 152 new empty codons to be assigned to ncAAs at will.

Codon Reassignment

In all organisms known the genetic code is degenerate, meaning that the number of codon is larger than the number of amino acids it needs to encode for; nonetheless, every codon is used and no free codons are available. A potential solution to the problem of expanding the amino acid pool for protein synthesis would be redefining the meaning of a fraction of those codons in a genome-wide fashion. While intuitively more straightforward, this strategy is significantly more challenging.

Since that the assignment of codons to amino acids is governed by aaRSs and their cognate tRNAs, reassignment of a particular triplet to a different amino acid means altering the natural tRNA(s) which decode(s) that particular codon. However, it is easy to imagine how the manipulations of the wild type set of tRNAs result in significant viability defects for the cells, due to interference with translation of the essential proteome. Alternatively, if the endogenous tRNA are not altered and, instead, an extra tRNA which can be aminoacylated with a ncAAs is introduced in the host, a partial reassignment of the codon is achieved which on the one hand limits the specificity of genetic code expansion, and on the other hand might show various degrees of toxicity associated with the production of dysfunctional proteins where the natural amino acids are replaced to some extent by ncAAs[30].

In a general sense, amber suppression represents a simple form of codon reassignment[70]. In this case, however, the meaning of the TAG codon is specified by the release factor 1 rather than by a tRNA. As such, amber suppression suffers from the limitations mentioned above in case of the addition of an additional tRNA without alteration of the endogenous pool, although to a very minor extent. In spite of it, the first attempts to fully reassign a codon were performed on the amber stop codon.

To obtain complete redefinition of the meaning of the TAG codon, two steps must be performed:

i) removal of all the undesired instances of this stop signal from the genome;

ii) deletion of the RF1 from the organism.

The first of these two steps is not trivial, considering the existence of over 300 annotated ORFs which are terminated by this stop codon[87]. In a first attempt to circumvent this problem, Mukai *et al.*[87] generated a BAC containing the 7 known essential genes which are terminated by an amber stop codon and edited them so that they would instead end with a TAA codon. In their experiment, they did not perform genomic manipulation of the amber-terminated genes in the bacterial host. The authors observed that, in the presence of this BAC, the otherwise essential *prfA* gene, coding for RF1, could be knocked out to produce a mutant strain called **RFzero**, but exclusively when an effective suppressor tRNA was present in the cell. In this strain, TAG was effectively redefined to both natural amino acids, like tyrosine, and to ncAAs like 3-iodo-L-tyrosine, however, growth was significantly

slowed down compared to the wild type strain containing the BAC. Johnson *et al*.[70] also investigated the possibility to knock out RF1 in the *E. coli* DMS42 strain and verified that, while normally essential in this strain, the *prfA* gene could indeed be deleted from the genome, but only after an alteration of the *prfB*, gene coding for RF2. In fact, the authors noted that common *E. coli* strains derived from the K-12 strain harbour a mutation in RF2 (Ala246Thr) which lowers its release activity on the UAA codon. By reverting this mutation, and by removing the natural regulatory premature opal codon within RF2[88], the authors could generated an RF1-knocked out strain, named **JX3.0**, where the UAG codon could be fully reassigned to an amino acid of choice. This strain also allowed multiple incorporation of a single new amino acid in response to multiple instances (up to 6) of the amber codon within a single CDS, without loss of efficiency.

A more systematic approach to try and edit all the instances of the amber codon was implemented by Farren Isaacs *et al.* by means of a genome editing method called multiplex automated genome engineering (MAGE)[89-91]. In this method, growing *E. coli* cells expressing the λ-red ssDNA-binding protein β are electroporated with a pool of ssDNA oligonucleotides containing the desired mutations, which are effectively used by the replication apparatus as primers for the synthesis of the lagging strand. Due to the parallel nature of the genome editing events, MAGE could be effectively used to induce the G to A mutation required to convert TAG codons into TAA codons. A subset of amber codons were mutated by this method in several different strains of *E. coli* whose genome was later combine by means of conjugative assembly genome engineering (CAGE), a method which takes advantage of conjugation, to generate the new C321 strain. This was further engineered by deleting the RF1 gene to produce the strain C321.ΔA, which displayed improved performance compared to the previously generated strains in terms of growth speed, but still significantly slower than the starting wild type strain.

Due to the constantly declining price for DNA synthesis, the chemical synthesis of recoded genomes has been investigated by other groups, too. In 2016, Wang *et al.*[92] investigated in a more methodic way which schemes for codon reassignment would be viable in *E. coli* strains in which a stretch of 20 kbp of genome, chosen to contain the highest density of essential genes, was replaced with synthetic recoded DNA. In their work, the authors designed a new λ-Red based recombineering technology, called **REXER**, in which the linear DNA substrate for recombination is generated *in vivo* by action of the CRISPR/Cas9 system and showed that it allowed efficient size-independent replacement of DNA pieces of up to 100 kbp. They then used REXER to edit the chosen 20 kbp of the *E. coli* genome in 8 different ways to show that some designs were not compatible with life, while others were viable. Later on, Fredens *et al.*[30] implemented REXER iteratively to assemble the first synthetic *E. coli*, named **Syn61**, where 2 serine codons and the amber codon were erased globally. In their work, they chemically synthesised an entire genome, recoded based on one of the reassignment schemes which

were found to be viable in their previous work, in fragments of around 10 kbp, and hierarchically assembled them into larger pieces to obtained 7 strains each containing between 0.5 to 1 Mbp of recoded genome. Finally, an individual strain containing a fully recoded genome was obtained by allowing the 7 intermediate strains to recombine following conjugation. The Syn61 strain doubles 1.6x slower than MDS42 in LB medium supplemented with glucose at 37°C, however, it is viable following deletions of the tRNAs which decode the two serine codons globally removed, allowing for the first time full reassignment of a sense codon in addition to the amber stop codon.

These works highlight how a significant effort is being made to create blank codons with the purpose of redefining the genetic code. This commitment, however, requires a parallel effort aimed at expanding the availability of mutually orthogonal aaRSs/tRNA pairs. In the next sessions, the current availability of pairs together with the difficulties connected to their discovery will be discussed.

Other Orthogonal Aminoacyl-tRNA Synthetase/tRNA Pairs

The TyrRS/$tRNA^{Tyr}$ from *M. jannaschii* and the PylRS/$tRNA^{Pyl}$ from *M. barkeri* or *M. mazei* have been extensively used for genetic code expansion due to their orthogonality and to the amenability of their active site to be engineered to accommodate substrates different from their wild type amino acid[54]. However, the use of such a restricted number of orthogonal pairs presents some limitations. In the first place, in spite of their versatility, both PylRS and TyrRS are mostly restricted to bulky and mostly hydrophobic amino acids which can be recognised by means of non-specific interaction[54]. Even when mutants are known to incorporate the ncAAs of interest, in order to incorporate multiple distinct ncAAs in different codons, multiple aaRS/tRNA pairs are necessary, each of which should be composed of:

(i) a tRNA which decodes a distinct codon;

(ii) an aaRS which accepts only one of the ncAAs of interest as a substrate;

(iii) an aaRS which aminoacylates only its cognate tRNA and not any of the ones of the other pairs or of the host's pairs.

In other words, multiple **mutually orthogonal** pairs are necessary. Lastly, even if PylRS does not recognise the anticodon of its cognate tRNA and hence is not affected by mutations in that portion of the tRNA, most of the other aaRS can tolerate minimal or no changes in the sequence of the anticodon of their cognate tRNA without the need of additional engineering. For all these reasons, availability of a large variety of mutually orthogonal aaRS/tRNA pairs represents an important aspect for the progression of the field, as it would increase the chances of successfully evolving a ncAA-RS for a

ncAA which differs chemically from what the currently available pairs can be engineered to accept. This consideration is especially true as genetic code expansion progresses towards the synthesis of polymers containing building blocks different from α-amino acids, which for example could be β-amino acids, α-hydroxy acids, α-thio acids etc.[28]

Following the initial identification of the *M. jannaschii* TyrRS/tRNA^{Tyr} pair and the failures of generating orthogonal pairs by evolution of the cellular aaRS/tRNAs[52], interest grew for the possibility to identify new pairs derived from organisms evolutionarily distant from the chosen host. One of the first of such attempts was the conversion of the yeast *S. cerevisiae* tRNA^{Asp} and its cognate *S. cerevisiae* AspRS into an amber suppressing pair[93]. The anticodon of tRNA^{Asp} represents a strong recognition element for the AspRS and the activity of the wild type enzyme towards the amber suppressor mutants of its tRNA^{Asp} is compromised. However, the mutation of a glutamic acid residue of the protein, which is responsible for mediating the interaction with the anticodon, to a lysine is able to partially rescue the aminoacylation of the tRNA. Pastrnak *et al.* observed that the AspRS^{E188K}/ $tRNA^{Asp}_{CUA}$ constitute an effective orthogonal amber suppressor pair in *E. coli*, but that the efficiency of suppression was very limited.

Soon after, a similar approach was used by Anderson *et al.* to identify an orthogonal LeuRS/tRNA^{Leu} pair[94]. By speculating that archaeal pairs might have naturally diverged enough to become orthogonal to their *E. coli* counterpart, the authors converted 5 known archaeal tRNA^{Leu} to amber suppressors by mutating their anticodon, taking advantage of the lack of interactions between LeuRS and the tRNA^{Leu} anticodon, then tested them alone or in combination to a limited selection of 6 archaeal LeuRS genes in an amber suppression assay. They observed that the LeuRS from *Methanobacterium thermoautotrophicum* and tRNA^{Leu} from *Halobacterium sp. NRC-1* can perform better than both the *Sc*-AspRS/ $tRNA^{Asp}_{CUA}$ and the *Mj*-TyrRS/ $tRNA^{Tyr}_{CUA}$ in amber suppression. Another pair composed of a tRNA and a synthetase from two different species was recently described as the *Archaeoglobus fulgidus* tRNA^{Ser}/*Methanosarcina mazei* SerRS[95]. Unfortunately, no later uses of these pairs for genetic code expansion have been reported yet.

Similarly, Chatterjee *et al.*[96] verified whether a combination among the ones formed by a group of 6 different ProRS and 6 mutant suppressors tRNA^{Pro} of archaeal origin could constitute an effective amber or quadruplet suppressor pair. Of those tested, they identified the *Pyrococcus horikoshii* ProRS as a good partner for the *Archaeoglobus fulgidus* tRNA^{Pro}.

An interesting alternative, which resulted in generation of two other pairs, has been undertaken by Santoro *et al.*[97] and by Anderson *et al.*[76], who generated *de novo* a tRNA^{Glu} and a tRNA^{Lys}, respectively, not based on a specific existing sequence, but from the consensus sequence of a set of archaeal tRNAs for the same isoacceptor class. Upon identification of orthogonal candidate tRNAs, the authors tested

their performances as triplet or quadruplet suppressors when combined with synthetases for their respective amino acids whose sequence was available. Anderson *et al.* additionally managed to engineer the LysRS from *P. horikoshii* to charge homoglutamine onto its cognate tRNA.

While several of the pairs mentioned above experienced a limited use in genetic code expansion, an interesting case is represented by the *E. coli* tryoptophanyl-tRNA synthetase/$tRNA^{Trp}$ pair. Italia et al. showed that this pair can be successfully used to expand the genetic code of mammalian cells, but also of *E. coli* itself[98], provided the genomic allele coding for this protein was previously replaced with the homologous orthogonal pair derived from *S. cerevisiae*[99]. Furthermore, it is worth noting that the *Ec*-TrpRS/$tRNA^{Trp}$ pair, in combination with the *Mj*-TyrRS/$tRNA^{Tyr}$ and the *Mb*-PylRS/$tRNA^{Pyl}$ were used to perform a triple incorporation of ncAA into a single protein in *E. coli*[100]. This result represents an important expansion on the previous work in which combinations of the pairs described above were used to perform dual incorporation of ncAAs[79, 101-105].

In all the above mentioned cases, orthogonality of the tRNA and or the synthetases was presumed based on consideration of evolutionary distance, then amber suppression was used to verify aminoacylation of the tRNA in the absence or presence of its cognate synthetase. In 2010 Yuan *et al.* [106] reported the identification of a subgroup of α-proteobacteria, among which is *Caulobacter crescentus*, that present a $tRNA^{His}$ missing a key identity element normally found in organisms across all kingdoms of life, namely an extra G at the 5'-end known as G(-1) which base pairs with the discriminator base C73. In light of this consideration, the authors verified that the *Cc*-HisRS does not aminoacylate *E. coli* tRNAs, while displaying activity towards its cognate tRNA, showing extreme specificity for its wild type anticodon. For this reason, orthogonality of the tRNA and aminoacylation by its cognate synthetase could not have been studied using amber suppression assays. Importantly, the authors did not test orthogonality of *Cc*-$tRNA^{His}$ *in vivo* in *E. coli*.

This study highlighted that orthogonality and activity of an aaRS/tRNA pair cannot always be assessed following alterations of the pair itself. In the case above mentioned, the *Cc*-HisRS would have been mistakenly considered inactive on its $tRNA^{His}$ if tested only in an amber suppression test. Instead, the enzyme is competent for aminoacylation in *E. coli*, but mutations of the tRNA anticodon abolish such interaction. Availability of this orthogonal pair, however, is valuable in itself in spite of its inability to perform as an amber suppressor, because it cannot be excluded that its activity on the mutant tRNA can be restored following engineering, as for the case of *S. cerevisiae* AspRS on its cognate $tRNA^{Asp}_{CUA}$, nor that the pair cannot direct incorporation of an amino acid of interest in response to a different codon, being it a sense, a quadruplet or a non-natural codon. In addition, the study highlighted how identifying tRNAs with new distinctive identity elements can successfully identify orthogonal pairs from other classes of organisms, while previous efforts had mostly inferred orthogonality based on generic considerations about phylogenetic distance. Furthermore, all the

studies mentioned were characterised by a rather narrow focus on few aaRS/tRNA combinations belonging to a specific isoacceptor class, and almost all of them make use suppression assays for the investigation. In the following paragraph I will discuss what makes the study of orthogonality of tRNA and synthetases difficult *in vivo* without relying on translation-based assays.

Testing Orthogonality of tRNAs and aaRSs

When expressed in a host organism, an orthogonal tRNA cannot be aminoacylated by any of the host's aaRSs, but it is substrate for its cognate synthetase. Similarly, an orthogonal synthetase exclusively aminoacylates its cognate tRNA, but does not aminoacylate any of the host's tRNAs. For any non-suppressor tRNA which has its native anticodon, easy-to-perform reporter assays cannot detect its aminoacylation status. In fact, as every sense codon in *E. coli* is read by the endogenous pairs, the production of the reporter cannot be correlated with the expression or aminoacylation of the potentially orthogonal tRNA. On the other hand, suppressor tRNAs are not natively present in the host, hence orthogonality of a suppressor tRNA can be deduced by the lack of production above background of a reporter interrupted by a stop or quadruplet codon in the presence of the tRNA of interest; production which should increase upon co-expression with its cognate synthetase.

When translation-based assays are not available, a different method needs to be used to verify the *in vivo* aminoacylation status of a tRNA. An important type of analysis on tRNAs was developed by Varshney *et al.* in 1991, when the authors observed that electrophoresis could be used to resolve the aminoacyl-tRNAs from their free counterparts under appropriate conditions[107]. In *E. coli*, tRNAs range in size from 74 to 95 bp[14], which equates to a molecular weight between ~24 and ~31 kDa. Conversely, the molecular weight of proteinogenic amino acids varies between 75 Da for glycine to 204 Da for tryptophan. As a result, the covalent attachment of an amino acid can at most increase the weight of a tRNA by <1%. In spite of this, the authors reported that running a tRNA sample on a urea polyacrylamide gel buffered at pH 5 using NaOAc was an effective way to resolve the two species over a run of about 20 cm. Given the poor electrophoretic mobility of tRNAs under those conditions, though, this PAGE require several hours to be completed. Furthermore, if more than one species is present in the sample, analysis of a specific tRNA of interest requires the use of northern blotting following the electrophoresis.

This method allows to verify *in vitro* whether aminoacylation occurred *in vivo* in a direct way without relying on a proxy. In spite of its surprising effectiveness, this approach is particularly laborious and time consuming to implement for the analysis of multiple different species. Other techniques developed later allowed to retrieve the sequences of the aminoacylated tRNA species by

using a combination of $NaIO_4$ oxidation, T4 ligation or poly-A addition and reverse transcription[108, 109]. Classical biochemical studies had shown that the strong oxidiser $NaIO_4$ is capable of reacting with the 3'-end of tRNAs, where the only vicinal diol moiety of the molecule is present, to form a dialdehyde[110]. However, when an amino acid is bound to the tRNA, one of the two hydroxyl groups at the 3'-end is esterified, hence the diol is protected from oxidation. As this reaction can be carried out at acidic pH when aminoacylation is stable, this oxidation permanently converts the aminoacylation status at a given time into a chemical fingerprint. Following oxidation, the amino acid can be cleaved off to expose a free 3'-end, amenable to enzymatic modifications such as adaptor ligation or poly-A polymerisation, while oxidised tRNAs cannot be modified. The samples can then be reverse transcribed and PCR amplified or sequenced to obtain information about aminoacylation. However, as tRNAs are highly structured and modified nucleic acids, reverse transcription is notoriously problematic on them[111], and furthermore orthogonal tRNAs get fully oxidised by $NaIO_4$, which means they are lost during the downstream processing if this method is used to measure orthogonality, such that the method would rely on negative data. Additionally, these methods require significant hands-on processing and have been primarily used to study the aminoacylation status of endogenous tRNA species.

Opposite to the case of tRNAs, orthogonality of the synthetases cannot be tested conclusively *in vivo* for any synthetase which recognises proteinogenic amino acids. In fact, given an aaRS for a particular isoacceptor class (e.g.: alanine), anomalous aminoacylation of the host's tRNAs for all the remaining classes with the incorrect amino acid (e.g.: alanine) can theoretically be investigated, but there is no experimental design to date which can reveal whether the synthetase under investigation can aminoacylate the endogenous tRNAs for the same isoacceptor class, because they would be always charged *in vivo* with that amino acid (alanine) regardless of whether the aaRS is orthogonal. Consequently, to circumvent this limitation most researchers have opted for *in vitro* aminoacylation assays, which are performed using the purified synthetase under investigation and a tRNA extract containing the deacylated host's tRNAs only, or the deacylated host's tRNAs together with the deacylated orthogonal tRNA[40, 93, 94, 97, 112]. In a hypothetical scenario, when provided with ATP and the appropriate amino acid, no aminoacylation reaction should occur when the synthetase is incubated only with the host's tRNA. However, this setup does not take into account that in the cell no synthetase is challenged with a high amount of deacylated tRNAs of many distinct isoacceptor classes, as in the cell a significant fraction of them are charged by their respective synthetases which have to compete to target their specific substrates. Consequently, this approach may potentially underestimate orthogonality for synthetases.

The considerations above highlight why identification of orthogonal pair has been attempted on a small scale by verification of the behaviour of a single or few combinations of tRNA/synthetase in the

host. This led us to consider which potential routes could be envisioned to establish a general way to identify new orthogonal pairs, independently of their isoacceptor class and of the identity of their anticodon.

Aim of the Project

In order to overcome the current limitations in availability of orthogonal aminoacyl-tRNA synthetase/tRNA pairs, we decided to develop a new approach which would take advantage of the knowledge of how *E. coli* aaRSs recognise their substrate but which could be applied on a larger scale and on pairs belonging to any isoacceptor class. We decided to set up a hierarchical approach which would focus on the identification of orthogonal tRNAs first, as they can be tested with greater ease, and later to the identification of a cognate active and orthogonal synthetase. Importantly, as alterations of a tRNA's anticodon can interfere both with its orthogonality and with its interaction with its cognate aaRS, we chose to perform our analysis on tRNAs with their native anticodon. It is important to notice that identifying orthogonal pairs this way does not ensure that these pairs can perform effectively in amber suppression, which is currently the most widely used technique for genetic code expansion, nonetheless, as the field progresses availability of new pairs will be required for applications beyond the suppression of stop codons.

In order to identify tRNA to test, we decided to rely on the natural divergence of these sequences among the multitude of living organisms. As next-generation sequencing becomes more and more applied, millions of new genomes become a valuable source for tRNA sequences, and for this reason we decided to generate an initial list of candidate tRNA starting from available databases data. As a key requirement for tRNA orthogonality is that, when introduced in *E. coli*, no aminoacylation should occur, we decided to filter the database developing a simple metric which would relate to the presence along the tRNA sequence of the identity elements for *E. coli* (*Ec*) aaRSs, under the assumption that sequences lacking these identity elements would have reduced likelihood of being aminoacylated. Having identified a list of candidates, we would experimentally test their behaviour in live cells. For the reasons discussed before, it was immediately clear that aminoacylation could not be tested relying on translation-based reporter assays, hence a new kind of assay had to be developed which could satisfy our experimental requirements.

While not being recognised by any of the *Ec*-aaRSs, orthogonal tRNAs must be charged by their cognate aaRS. For tRNAs which could not be aminoacylated *in vivo* in *E. coli*, we decided to rely on

genomic data to identify, in the genome from which the tRNA was identified, the cognate aaRS for the same isoacceptor class as the tRNA under investigation (e.g.: AspRS for a $tRNA^{Asp}$ etc.). Upon co-expression of the two components of the pairs together in the same cells, aminoacylation of the heterologous tRNA implies that the synthetase is correctly expressed in a catalytically proficient way and that it can effectively recognise its tRNA. However, this test is not sufficient to make sure that the heterologous synthetase cannot mis-aminoacylate any of the other endogenous *Ec*-tRNAs. We

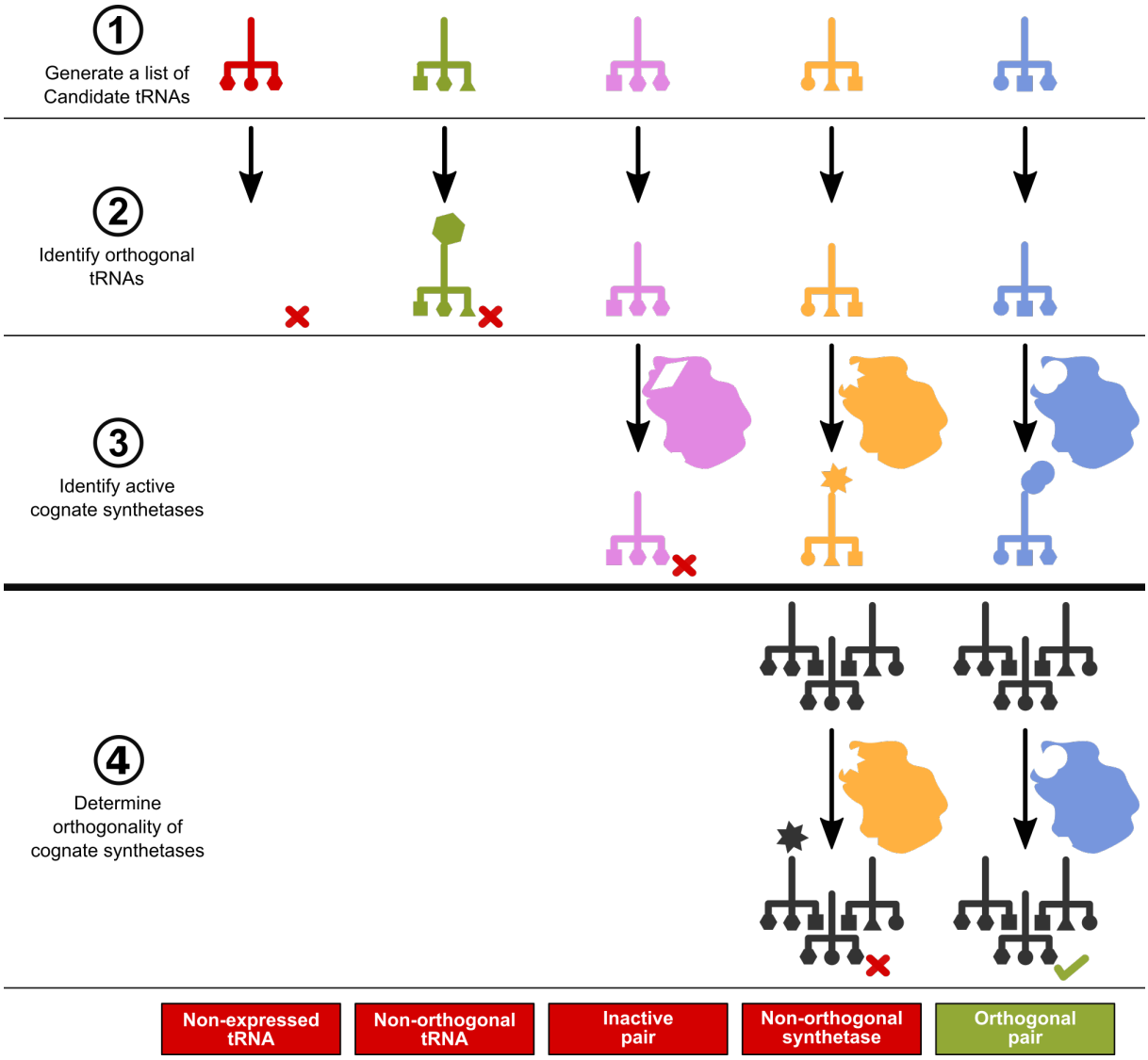


***Figure 1.3**: Schematic representation of the pipeline developed for the identification of aminoacyl-tRNA synthetase/tRNA pairs identified in living species and orthogonal in E. coli.* **(1)** *Generation of a list of candidate orthogonal tRNAs, selected among the ones annotated in tRNA databases, with a higher likelihood of being orthogonal in E. coli based on preliminary considerations.* **(2)** *expression of the candidate tRNAs in E. coli to verify which ones can be detected in cytosolic extracts but which cannot be aminoacylated by E. coli aaRSs.* **(3)** *Co-expression of the tRNA and cognate aaRS from the same organism as the tRNA in E. coli to identify active pairs in which aminoacylation of the tRNA is dependent on the presence of its cognate synthetase.* **(4)** *Verification of the orthogonality of the exogenous aaRSs with respect to the endogenous E. coli tRNAs to confirm orthogonality of the pair as a whole.*

decided to test orthogonality for these enzymes *in vitro*, in spite of the limitations of this kind of assays.

In summary, the problem of identifying orthogonal pair was divided into a sequence of steps, each of which focused on:

i) generation of a list of candidate orthogonal tRNAs;

ii) characterisation of the interaction between the candidate tRNAs and the *E. coli* aaRSs;

iii) characterisation of the interaction between the orthogonal tRNAs and their cognate aaRS;

iv) characterisation of the interaction between active aaRSs and *E. coli* tRNAs. (**Figure 1.3**).

In the next chapter I will describe how we implemented the strategy described herein to screen a large number of tRNAs and synthetases and how this allowed us to identify new orthogonal pairs. I will then describe how I attempted to engineer these pairs for the particular use of amber suppression for genetic code expansion and how this engineering process improved the activity and orthogonality of the pairs.

Chapter II – Identification of Orthogonal Aminoacyl-tRNA Synthetase/tRNA Pairs

Computational Analysis

As next-generation sequencing technology progresses and our capability to read entire genomes becomes faster and cheaper, large collections of data are now publicly available in the form of annotated databases. In 2011 a group of researchers from Japan led by Takashi Abe produced a work in which they reported the creation of a database, tRNA-DB-CE, containing the sequence information of the tRNAs present in these deposited genomes[26]. The authors decided to rely on three distinct searching algorithms, tRNAscan-SE[113], ARAGORN[114] and tRNAfinder[115] to maximise the accuracy of the identification of tRNAs and furthermore they verified individually all the instances in which the

predictions from these three softwares differed. Their database stores information about primary and secondary structures, anticodon, species and genome of origin for several millions of tRNAs from bacteria, archaea, phages, chloroplasts and metagenomes.

In our study we decided to exploit the variability generated by the natural divergence among species to identify tRNAs which lack the distinctive features recognised by *E. coli* aaRSs, which we hypothesised would make them more likely to be orthogonal in this host. In order to do this, we used the information contained in the tRNA-DB-CE as our starting point. The computational analysis which will be described in the following sections was performed together with Dr. Stephen Fried.

Reliable tRNAs for every isoacceptor class from complete or drafted bacterial and archaeal genomes, together with those from phages and chloroplasts, were downloaded and split based on their isoacceptor class. The database contained 2,799,231 entries as of March 2017 including tRNAs for selenocysteine and natural suppressor tRNAs. Importantly, as the isoacceptor class is assigned to each entry based on the sequence of its anticodon, special attention had to be paid when considering the $tRNA^{Ile}$ with anticodon CAU. In fact, while ATG is the methionine codon in the most widespread genetic code, the ATA codes for isoleucine, so that the cells would require a $tRNA^{Ile}$ with anticodon UAU. However, as this last anticodon would base pair with the codon ATA but also wobble with the methionine anticodon ATG, cells have developed a remarkable post-transcriptional modification to $tRNA^{Ile}_{CAU}$ which converts C34 to lysidine, an unusual nucleotide which base pairs with A without wobbling with G due to steric clashes. As a result, $tRNA^{Ile}_{CAU}$ annotated as $tRNA^{Met}$ were manually corrected to $tRNA^{Ile}$. In addition, $tRNA^{Met}$ were divided between initiators and elongators, and only the latter were used for the analysis. Following this initial processing of the data, all the sequences were aligned to a standard canonical tRNA archetype, as described below.

tRNAs Alignment to Canonical Form

E. coli genome contains 86 tRNA genes of which 48 distinct sequences used to decode for the 20 canonical amino acid and selenocysteine14. In spite of the general diversity in their primary sequence, all of them share some features:

1. The acceptor stem is 7 base pair long for all canonical amino acids, with the selenocysteinyl-tRNA being the only exception with a distinctive 8 base pairs long acceptor stem;
2. base 8 is always U and does not base pair, and neither does base 9;
3. the sequence GG present in the D loop is assigned always to canonical positions 18 and 19, in

order to do so occasionally positions 17a, 20a and 20b are assigned;

4. the anticodon stem is 5 base pairs long, the anticodon loop is 7 nucleotides long;
5. positions 54→56 are assigned to the sequence TΨC which is conserved for all of them; the TΨC stem is 5 base pairs long and the loop is 7 nucleotides long.
6. positions 74→76 have the conserved sequence CCA across all kingdoms of life.

Given the conservation of these general features, tRNAs can be aligned to a canonical form. As the variable loop occasionally contains only 4 nucleotides, in the alignment shown position 46 might be not assigned. To proceed with our analysis, we decided to align all tRNAs from the database to this canonical form. Notably, positions 74→76 are always represented by the nucleotides C, C and A even if the tRNA gene has a different sequence for those nucleotides, as many organism, among which is *E. coli*, possess a quality control apparatus able to add this 5'-CCA-3' tail to tRNAs terminating with a different sequence[116]. As a result, these positions are not informative and a 73 nt alignment was generated.

The conversion to this canonical form is mostly straightforward given the secondary structure of a tRNA, but in some circumstances decisions were taken discretionally:

- for tRNAs with acceptor stem 8 nucleotides long, the first 7 base pairs were assigned to positions 1→7/66→72, considering that the only *E. coli* tRNA with such feature, $tRNA^{SeC}$, aligns in such manner to the $tRNA^{Ser}$, with which it shares the recognition by the SerRS;
- if the D arm is formed of 3 base pairs, positions 13 and 22 are assigned to the first and last nucleotides of the D loop;
- in *E. coli* the D loop presents some degree of variability, but every tRNA presents two distinctive Gs which are always assigned to positions 18 and 19. As the loop starts at canonical position 14 and ends at canonical position 21, to account for the difference in length of this loop sometimes it is necessary to assign the additional positions 17a, 20a or 20b or leave some positions unassigned in *E. coli* tRNAs. Given the higher digree of variability found among the tRNAs in the database, and given the frequent lack of G18 and G19, the alignment for this structural element was manually generated as shown in **Chapter V – Appendix: D Loop Alignment Table**. The extra positions 17a, 20a and 20b are not used in the 73 nt alignement;
- as the variable loop can have a very broad range of lengths, the positions 44→48 were assigned in the following order: 44→48→45→47→46. For loops longer than 5 nucleotides, the first two positions are numbered 44 and 45, while the last three are numbered 46→48.

Following the above listed rules, all the tRNAs were converted into a multiple alignment. As an

example, let us consider a tRNA with the following sequence, corresponding to the $tRNA^{Arg}$ from *Aquifex aeolicus VF5* with accession number C000013 in tRNA-DB-CE:

```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 20a 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76
GGGCCCGTAGCTCAGTTGGATAGAGCAGGGGATTTCTAATCCCCAGGTCGGGGGTTCGAGTCCCTCCGGGCCCACCA
(((((((..((((.........)))).(((((.......))))).....(((((.......)))))))))))....
```

The top line represents the canonical numbering scheme, the bottom line is a conventional way of representing secondary structures of nucleic acids when no knots are presents. In particular, nucleotides in open brackets base pair with nucleotides in close brackets in the order in which the brackets appear (e.g.: the first closed bracket base pairs with the last open bracket). Unpaired nucleotides are represented by dots.

In the case above, the D loop has a length of 9, so that position 20a needs to be specified. Given the alignment, the 73 nt form is obtained by removing positions 20a and 74→76 to obtain

```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73
GGGCCCGTAGCTCAGTTGGAAGAGCAGGGGATTTCTAATCCCCAGGTCGGGGGTTCGAGTCCCTCCGGGCCCA
(((((((..((((.........)))).(((((.......))))).....(((((.......)))))))))))).
```

After having performed this type of alignment, the 20 isoacceptor classes were highly redundant, both because the same tRNA can be annotated in multiple organisms, in multiple strains or in multiple genome assemblies of the same organism, and because distinct tRNAs can have the same alignment by compression of the sequence in the D and variable loops. Sequences with the same 73 nt alignment were compressed together, then each unique sequence was evaluated by a scoring algorithm, as described next, to highlight which one is more likely to be orthogonal.

In an attempt to develop a filter which could be applied easily, we decided to undertake a simplified approach where each tRNA is scored for its potential to interact with each of *E. coli* aaRSs simply counting how many of its nucleotides at the positions which are recognised as identity element by that particular aaRS match the nucleotides which are found on the aaRSs' endogenous tRNA substrate. As an example, let us consider again the $tRNA^{Arg}$ from *Aquifex aeolicus VF5*:

```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73
GGGCCCGTAGCTCAGTTGGAAGAGCAGGGGATTTCTAATCCCCAGGTCGGGGGTTCGAGTCCCTCCGGGCCCA
(((((((..((((.........)))).(((((.......))))).....(((((.......)))))))))))).
```

Let us now try to observe how it compares to the *Ec*-$tRNA^{Trp}$:

```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73
AGGGGCGTAGTTCAATTGGTAGAGCACCGGTCTCCAAAACCGGGGTGTTGGGAGTTCGAGTCTCTCCGCCCCTG
(((((((..((((.........)))).(((((.......))))).....(((((.......)))))))))))).
```

As mentioned before, in a simplified model we can hypothesise that *Ec*-TrpRS only tests which nucleotides are present at positions 1→3, 34→36 and 70→73 (highlighted in red for both tRNAs).

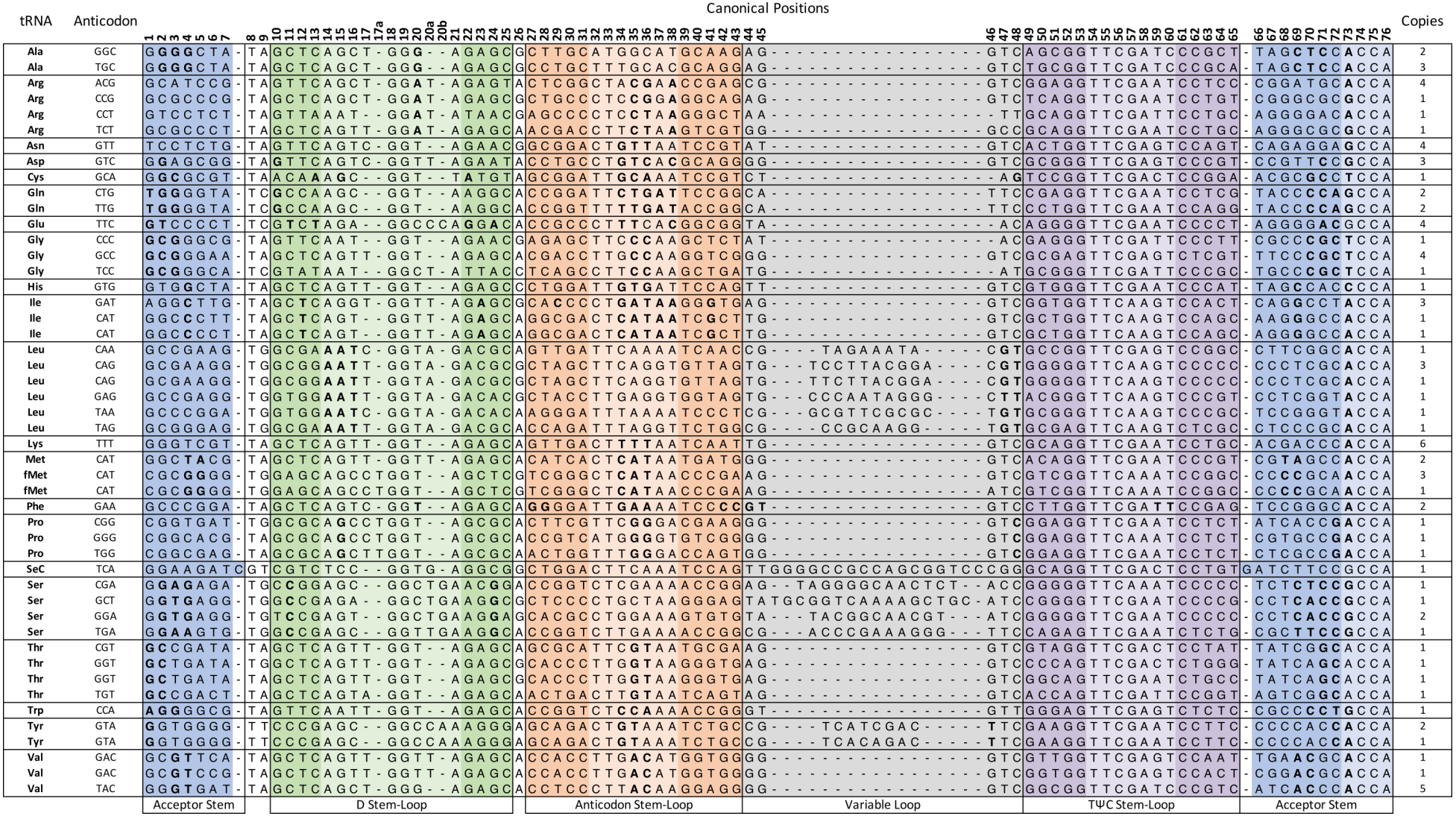


***Figure 2.1**: structural alignment of each of the 48 distinct E. coli tRNAs sequences to the canonical numbering scheme (top row), listed for isoacceptor class (first row). Several tRNA genes are contained in the genome in multiple copies, as indicated in the last column. The elements of the secondary structure (bottom row) are colour-coded: acceptor stem (blue), D arm (green), anticodon arm (orange) variable loop (gray), TΨC arm (purple).*

The two tRNAs have the same nucleotides at positions 2, 3, 35, 70 and 71, but differ at positions 1, 34, 36, 72 and 73. If each matching position is scored +1 and each different position is scored -1, the total score of *Aquifex aeolicus VF5* $tRNA^{Arg}$ for the identity elements of *Ec*-TrpRS is calculated as the average [5(+1)+5(-1)]/10=0.

Scores thus calculated are the mean of a set of numbers which can only be -1 or +1 and for this reason each score is contained within these boundaries. Analogously, *Aquifex aeolicus VF5* $tRNA^{Arg}$ can be compared to each of the 47 *E. coli* tRNAs for the 20 natural amino acid to verify how similar it is to known substrates for the *E. coli* aaRSs, generating 47 scores. For synthetases which can aminoacylate multiple isoacceptor tRNAs, as in the case of LeuRS or SerRS etc., the comparison was performed between the tRNA under investigation and each of the *E. coli* isoacceptors individually, then the scores were averaged to obtain a single score, so that eventually each tRNA received 20 scores, one for each aaRS.

This approach evaluates every identity element with the same weight, while it is likely that not all interactions contribute equally to the binding energy. As an example, the presence of a G in position 34 or 35 of the anticodon might completely prevent binding of a tRNA to the TrpRS due to stearic clash of a larger purine in place of a pyrimidine; however characterisation of these types of interactions would require extensive experimental validation and cannot be effectively modelled *a priori*. We hence decided to verify if this simple scoring system could be used to model the interaction between tRNAs and aaRSs and in order to achieve this validation we verified how the set of *E. coli* tRNAs themselves, which are naturally mutually orthogonal, would be evaluated by the metric described.

E. coli tRNAs Scoring

In the scoring system described above, the endogenous *E. coli* tRNAs are effectively used as model substrates for their synthetases, while the aim of the procedure is to estimate whether a particular tRNA of interest would be a substrate for the *E. coli* aaRSs. In order to validate how the score for a given aaRS correlates with the likelihood of recognition and aminoacylation by that aaRS, we needed a training set of tRNAs whose interaction with the *E. coli* enzymes is known. In order to refine the scoring to achieve improved predictive power within the limits of the simplifications applied, a large training set is required. However, we reasoned that the largest set of sequences whose interactions with *E. coli* synthetases is experimentally validated was constituted by the *E. coli* tRNAs themselves, with the limited addition of few orthogonal tRNAs which are known not to interact with any of them. In fact, as a requirement for the fidelity of protein synthesis, each of the endogenous tRNAs must be

aminoacylated exclusively by its cognate aaRS while being orthogonal to the other 19 present in the cell.

As the *E. coli* tRNAs are present in the database, we retrieved the scores assigned to them by the algorithm described (**Figure 2.2a**). For each isoacceptor class containing only one tRNA, the score attributed to that tRNA for the corresponding synthetase was equal to +1 by definition. When multiple isoacceptors are present, the score can still be +1 if no variation is present in any of the isoacceptors at the identity elements, as in the cases for glycine, proline, threonine, tyrosine or valine. For other classes the score for their cognate aaRS was lower than +1, nonetheless in most cases it was the highest among the scores for all synthetases, indicating a strong correlation between the scores and the aminoacylation *in vivo*. There are, however, some exceptions to this observation. For example, the *Ec*− $tRNA^{Arg}_{CCT}$ has a higher score for the IleRS and for the LeuRS, which would suggest an interaction which is not known to occur. Given the simplicity of the model, however, a number of false positive and a number of false negative predictions are expected to occur. In fact, it is important to notice that identity elements are often characterised as the positions along a tRNA which, when mutated, abolish aminoacylation by the cognate aaRS. In this context, identity elements behave as nucleotides *necessary* for aminoacylation. When studing aminoacylation of a tRNA by a non-cognate aaRS, on the contrary, it is crucial to understand if a given set of nucleotides is *sufficient* to lead to recognition of that tRNA by the aaRS. For this reason, it is easy to hypothesise that our knowledge on the identity elements which are sufficient for mis-aminoacylation is incomplete, resulting in a loss of predictive power by the algorithm. In spite of these considerations, nonetheless, higher scores were calculated for the cognate synthetases of each tRNA, while the largest portion of the score matrix was composed of entries ≤0 (**Figure 2.2a**).

The score matrix for individual isoacceptors did not identify a specific threshold above which a score unambiguously predicted a successful interaction and below which aminoacylation could not occur. In order to account for the variation within a given isoacceptor class, we calculated the average scores that tRNAs from the same isoacceptor class were given for each of the *E. coli* aaRSs (**Figure 2.2b**). In

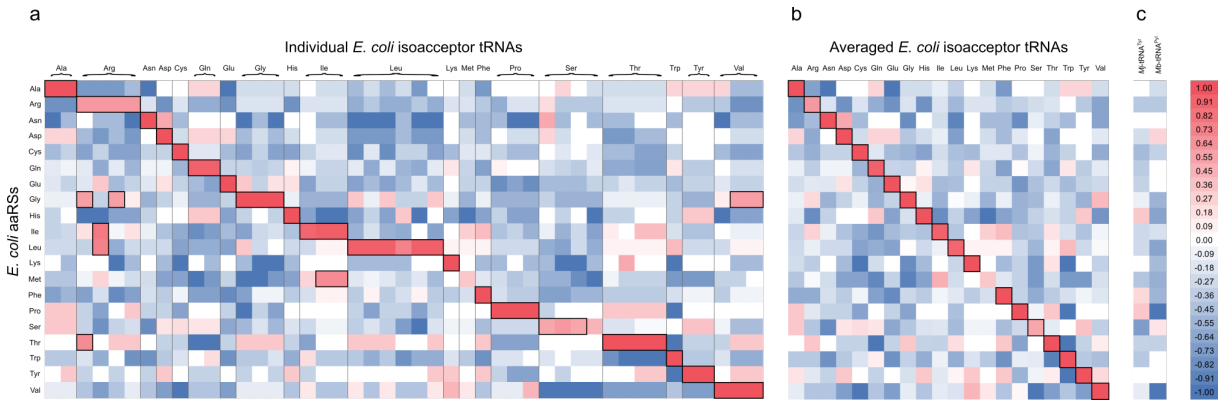


***Figure 2.2**: **a)** Scores for each of E. coli isoacceptor tRNAs individually. **b)** Average scores of isoacceptor tRNAs. **c)** Scores for the two known orthogonal tRNAs Mj-tRNA^{Tyr} and Mb-tRNA^{Pyl}. Scores >0.50 are highlighted with a box.*

this new matrix, we observed that values of >0.50 only occurred along the diagonal, while orthogonality between a tRNA and an aaRS effectively correlated with a score of ≤0.50. Notably, the elongator $tRNA^{Met}$ scores only 0.33 for the MetRS, for which it is a known substrate. This is due to the divergence in sequence between the elongator and the initiator tRNAs, however a lower score does not imply orthogonality in this case.

In summary, the matrix of the average scores for each isoacceptor class suggested that a score of >0.50 is a good indicator for interactions, while a score of ≤0.50 most of the times correlated with orthogonality. Clearly, this metric is not completely predictive, however it is important to highlight that the primary goal of our investigation was to develop a general pipeline for the identification of orthogonal pairs, not the validation of an algorithm with high predictive power. In particular, we intended to perform a simple automatic filtering procedure which would exclude from a large starting dataset the tRNAs which are very similar to the endogenous *E. coli* tRNAs, hence less likely to be orthogonal, while highlighting entries whose unusual characteristics might have resulted in a higher chance of being orthogonal. Furthermore, the experimental characterisation of a larger number of tRNAs is an important step to calibrate a more predictive metric. Hence, we chose to verify if a threshold of 0.50 would correctly identify the *Mj*-$tRNA^{Tyr}$ and *Mb*-$tRNA^{Pyl}$ as orthogonal (**Figure 2.2c**). While the former tRNA sequence is contained in the dataset, the latter, being a natural amber suppressor tRNA, was not included, hence it was manually aligned taking into account the unusual characteristics of its secondary structure, then scored by the algorithm. Importantly, none of the scores for these two known orthogonal tRNAs were >0.50.

Overall, the scores generated by the procedure described seemed to recapitulate quite effectively the interaction network among *Ec*-tRNAs and *Ec*-aaRSs. Noticeably, effective interactions which result in aminoacylation within our dataset are assigned positive scores almost always >0.50, while in some instances successful interactions occur when the score is ≤0.50 but >0.00. Conversely, interactions between a tRNA and a non-cognate synthetase are largely characterised by scores ≤0.50, even though in some specific cases a high score did not correspond to incorrect recognition.

Given that the sequences available in the database were several orders of magnitude higher than the number of tRNAs which could be experimentally tested, and in light of the observation stated above, we decided to undertake a conservative approach in the interpretation of the scoring results. In particular, we

i) filtered out of our dataset all tRNAs with a score of >0.00 for the *Ec*-aaRS for the same isoacceptor class;

ii) classified the remaining tRNAs in two groups: sequences for which any of the other scores were >0.50 were classified as Tier 1 candidates, while tRNAs for which all the other scores were ≤0.50 were classified as Tier 2 candidates.

This process produced a list of candidate tRNAs for each isoacceptor class whose size was significantly reduced compared to the initial number of distinct entries in our dataset (**Figure 2.3**). Among these, we chose a subset of about 10 to 30 tRNAs per isoacceptor class, mostly among the Tier 2 candidates but including some from Tier 1, which we experimentally tested for orthogonality in *E. coli*.

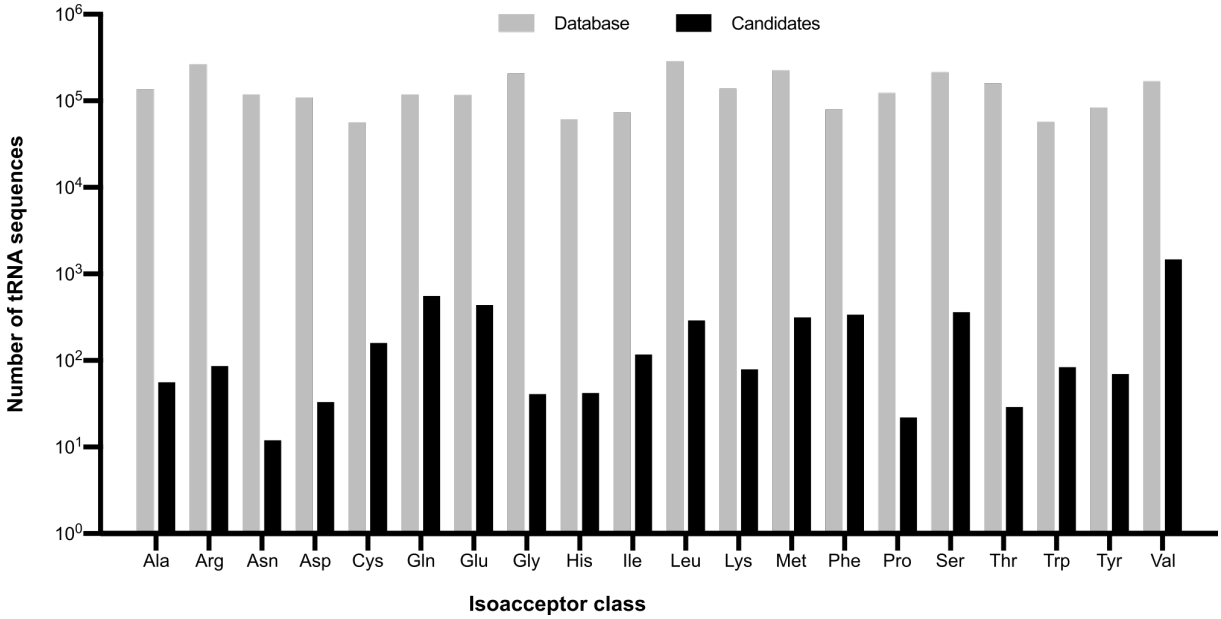


***Figure 2.3**: Number of distinct entries, sorted based on their isoacceptor class and displayed on a logarithmic scale, downloaded from the database and composing the initial dataset (grey bars), and sequences selected after the application of the filter described in the text (black bars). The filter effectively reduced the number of entries of 2 to 3 orders of magnitude.*

An *in vitro* Assay to Test Aminoacylation: tREX

The development of the scoring system described before tackled the first of the goals of the project (**Figure 1.3 (1)**), i.e.: generating a list of candidate orthogonal tRNAs containing sequence more likely to be orthogonal. In order to successfully identify orthogonal pairs, however, experimental validation was required to verify whether those tRNAs would be expressed and/or aminoacylated when introduced into our host of choice, *E. coli*. While methods to test the aminoacylation status of a tRNA of interest were available (see **Testing Orthogonality of tRNAs and aaRSs**), all of them were developed to test moderate number of samples and were laborious and time-consuming, so that none of those were appropriate to efficiently screen a few hundred tRNAs, as we had committed to do. In order to overcome the limitations which those techniques had, I wanted to create an assay which would satisfy three conditions:

i) to be independent of the anticodon of the tRNA under analysis;

ii) to be independent of the nature of the amino acid bound to the tRNA of interest;

iii) to be easily scalable to allow investigation of multiple tRNA in parallel.

To achieve such results, a new assay would have to rely on few steps of manipulation of the samples and on a non labour-intensive detection method. Due to its widespread use, I started by analysing what factors were limiting in the application of acidic urea PAGE to our project. This technique was completely independent of the anticodon of the tRNA of interest – hence condition *i*) was satisfied – and requires essentially no manipulation of the tRNA samples after extraction, because the difference in electrophoretic mobility between the aminoacylated and free tRNA is given by the amino acid itself, hence they can be immediately run on the gel. However, this direct link between amino acid and electrophoretic mobility implies that distinct amino acids cause variable retardation to the migration speed – thus not satisfying condition *ii*). Furthermore, this difference can only be observed after a migration of several tens of centimetres, which in those conditions require several hours. Additionally, visualisation of the specific tRNA of interest in the complex mixture of cellular tRNAs relies on northern blotting. As a consequence, the manipulation of the sample following the electrophoresis was quite significant, as each tRNA needed to be assayed by a different probe on a different membrane, thus not satisfying condition *iii*). These practical considerations excluded the possibility to use this technique for our purpose. However, I reasoned that converting the aminoacylation status of a tRNA into an alteration of its electrophoretic mobility could be a fruitful way to approach the problem. I decided to explore variations of this method in which:

a) detection would not require additional steps following electrophoresis;

b) the difference in electrophoretic mobility between the aminoacylated and free species would be generated indirectly and would be larger than the one caused directly by the presence of the amino acid, hence eliminating the need of running lengthy gels.

To implement variation *a*), I decided to alter the design of the probes used for northern, which are normally biotinylated, and to label them using a fluorophore instead. In fact, if hybridisation is performed *before* electrophoresis, this variation allows to directly visualise the specific target among all the cellular tRNAs by means of *in gel* fluorescence. Clearly, this approach cannot be undertaken in the context of acidic urea PAGE, as the binding of a probe to the tRNA of interest would greatly reduce the retardation caused by the amino acid. As nucleic acid are commonly stained with green-fluorescent dyes (e.g.: SYBR Gold) which absorb blue light around 488 nm and emit green light around 520 nm, probes labelled using the cyanine 5 (Cy-5), which absorbs around 635 nm and emits around 670 nm allows multi-colour imaging without any spectral overlap among different fluorophores.

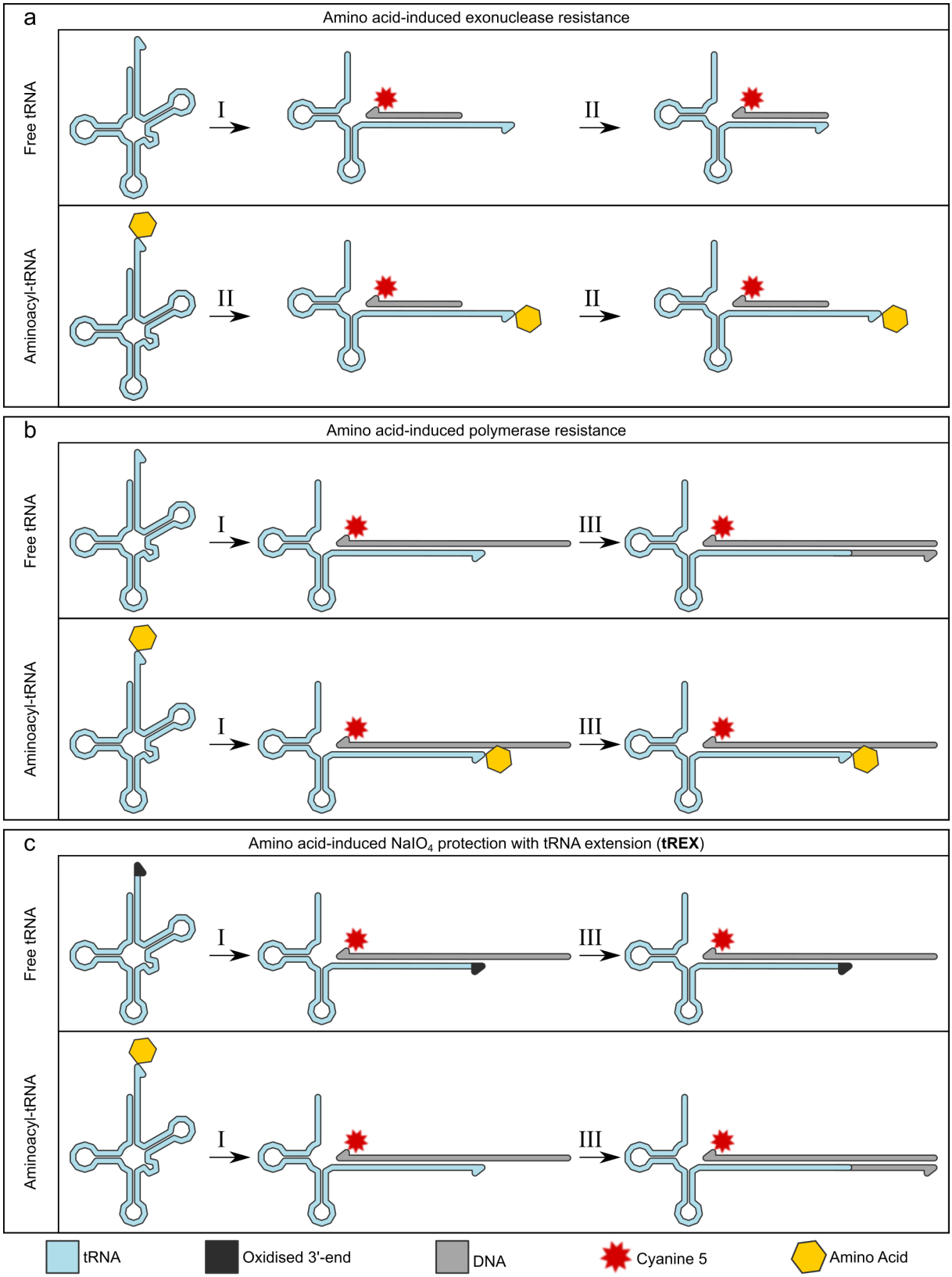


***Figure 2.4**: Designs for in vitro aminoacylation assays. I: probe hybridisation. II: RNAse T treatment. III: DNA polymerase treatment.*

However, envisioning an experimental design that would implement the variation *b*) was more challenging. In a first attempt, I hypothesised the use of a 3'-to-5' ssRNA-specific exonuclease to reduce the size of free tRNAs, under the working hypothesis that aminoacylation could serve as a protecting group to prevent the enzyme from digesting aminoacylated tRNAs (**Figure 2.4a**). This would result in an increase in mobility for the free species dependent on the degree of digestion of its 3'-end. Unfortunately, this design is limited by the fact that the only available RNAse presenting the required characteristics is RNAse T, an enzyme naturally involved in trimming of the 3'-end of tRNAs. This nuclease has evolved to stall when the sequence 5'-CC-3' occupies its active site, which always occurs immediately after cleaving of the universally conserved A at the very 3'-end of any tRNA. As a consequence, this method only allows removal of a single nucleotide from tRNAs, which is detectable on gels, but does not constitute the significant improvement I desired.

Alternatively, I hypothesised that extension of the free 3'-end of non-aminoacylated tRNAs could be more versatile to induce a tunable variation in electrophoretic mobility. In this idea, the DNA probe would be annealed to the 3'-end of its target tRNA generating a protruding 5'-end of ssDNA and a recessed 3'-end of RNA (**Figure 2.4b**). As DNA polymerases, like the Klenow exonuclease-deficient fragment of *E. coli* DNA polymerase I, can easily use RNA as a primer, in this design the probe also serves as a template for the extension of the tRNA with extra nucleotides. In this context, it is easy to imagine how aminoacylation would prevent extension, hence generating an electrophoretic retardation which can be altered at will, since the template can be arbitrarily chosen of any length and sequence.

While this alternative was significantly more promising than the previous one, it presented a key limitation. In fact using the aminoacylation as a protecting group can be effective exclusively under those conditions which preserve it, i.e.: at pH 5. At this pH, many enzymes are not catalytically competent, and in particular the Klenow fragment of DNA polymerases I did not display any activity. Conversely, at a more favourable pH in which the polymerases are active, aminoacylation is labile and full extension of the tRNA of interest is observed (see below).

A key contribution which allowed to simply modify the last design and make it functional came from the well established $NaIO_4$ oxidation of tRNAs, which had been previously used extensively for example in the OXOPAP method[109]. In fact, while I had experimentally observed extension of the tRNA primer over a DNA probe/template, in my experiment the amino acid's role as protecting group was compromised by the liability of the aminoacyl-tRNA ester bond. Conversely, protection of the 3'-end of aminoacyl-tRNA from $NaIO_4$ oxidation was well established and known to be both highly effective and highly efficient, as it can be carried out at acidic pH to completion in about 1h on ice. I hence modified my idea by performing oxidation of the tRNAs immediately following extraction and before hybridisation with the probe and extension by the DNA polymerase (**Figure 2.4c**). It must

be noticed that this procedure inverted the alteration in electrophoretic mobility compared to its original design (**Figure 2.4b**), meaning that the high molecular-weight band corresponds to aminoacylated species whose 3'-end had been protected from oxidation.

To verify the efficacy of the procedure, I extracted total tRNAs from *E. coli* cells, used as a negative control, and from *E. coli* cells expressing the *M. mazei* PylRS and a cognate $tRNA^{Pyl}$ from *M. barkeri*. I designed a DNA probe complementary to the 3'-end of the tRNA and that invades its variable loop, T arm and to the second strand of the acceptor stem (**Figure 2.4c**). Furthermore, the probe had a poly-A tail at its 5'-end which could be used by the DNA polymerase as a template, and a cyanine 5 (Cy-5) covalently bound to its 3'-end.

Annealing of the probe to the *E. coli* tRNA extracts produced a high-molecular weight band only in the presence of $tRNA^{Pyl}$ (**Figure 2.5** lanes 3,7). Incubation of the tRNA/DNA heteroduplex with the

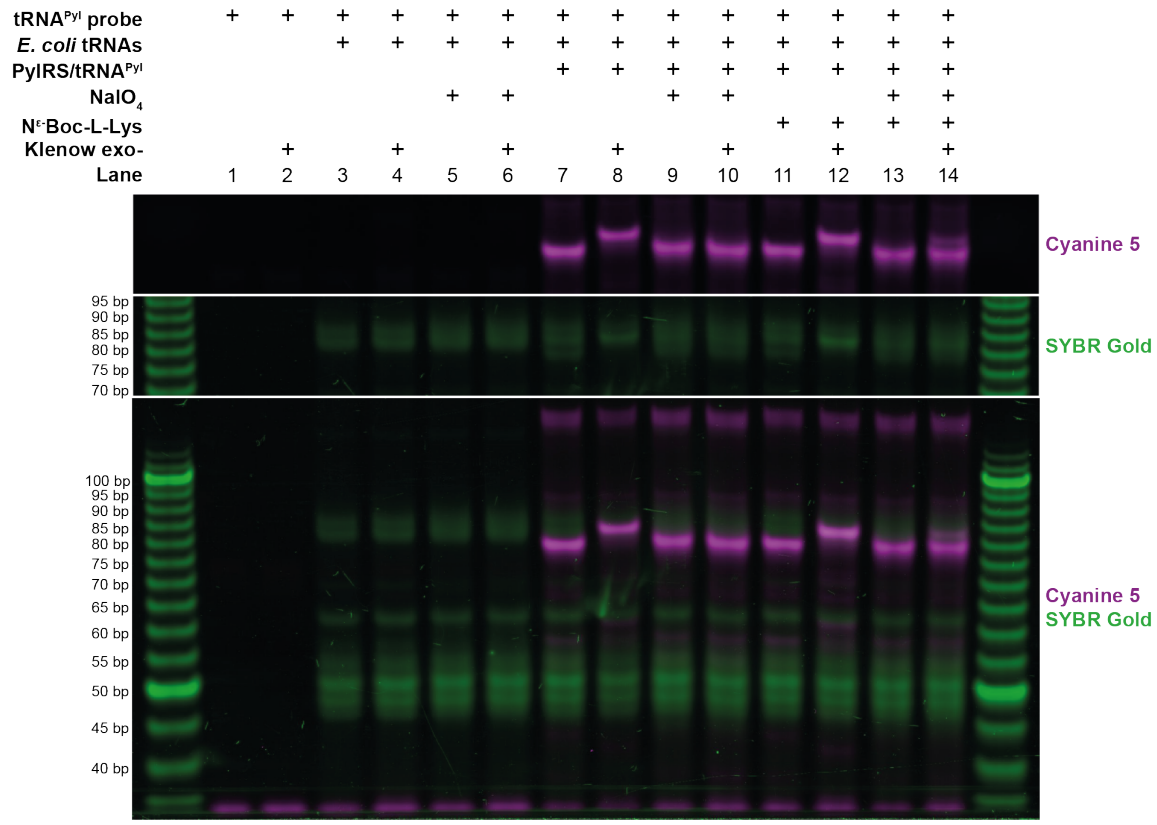


***Figure 2.5**: Experimental validation of tREX using PylRS/$tRNA^{Pyl}$ as a model system. In the absence of any tRNA extract (lanes 1-2) or in the presence of tRNA extracts not containing $tRNA^{Pyl}$, the DNA probe for the detection of $tRNA^{Pyl}$ did not give non-specific signals, independently of the sample processing. In the presence of $tRNA^{Pyl}$, the probe generated a band in the gel (lane 7) which could be extended by the DNA polymerase to reduce its electrophoretic mobility (lane 8). The extension by the polymerase was inhibited by oxidation of the 3'-end of the $tRNA^{Pyl}$ by $NaIO_4$ (lane 10). Aminoacylation of $tRNA^{Pyl}$ by PylRS, induced by the addition of N^{ε}-boc-L-lysine to the growth medium of the cells from which tRNAs are extracted, did not prevent extension of the tRNA/probe heteroduplex by the DNA polymerase (lane 12). However, oxidation of the aminoacylated $tRNA^{Pyl}$ prior to the enzymatic extension by the polymerase allowed the resolution of two bands on PAGE (lane 14).*

exonuclease-deficient Klenow fragment of DNA polymerase I led to an decrease in the mobility of the Cy-5 labelled species, which was easily visible in a gel of standard length (~8 cm) (**Figure 2.5** lanes 7, 8). As noted before, in the absence of $NaIO_4$ oxidation full extension of the tRNAs was observed regardless of whether the cells were grown in presence or absence of N^{ε}-Boc-L-lysine (BocK), a substrate for PylRS which is charged on $tRNA^{Pyl}$ (**Figure 2.5** lanes 8,12). This observation proves how aminoacylation alone is not sufficient to protect the tRNA from extension by the DNA polymerase under the conditions required for the assay. Importantly, treating the tRNA extract with $NaIO_4$ before the hybridisation procedure completely prevents free tRNAs from being extended (**Figure 2.5** lanes 10), consistent with quantitative oxidation of $tRNA^{Pyl}$. Instead, when the tRNA extract from *E. coli* expressing PylRS and $tRNA^{Pyl}$ and grown in presence of 1 mM N^{ε}-Boc-L-lysine was oxidised, hybridised to the DNA probe and incubated with the DNA polymerase, two bands were resolved on the gel (**Figure 2.5** lanes 14), which migrated at the same height in the gel as the native and extended $tRNA^{Pyl}$, respectively. These experiment confirms that my new assay, which exploits the probe-templated extension by a DNA polymerase of a candidate tRNA following $NaIO_4$ oxidation, can be used to read *in vitro* the aminoacylation status *in vivo* of a tRNA of interest. Given its design, I named this new technique **tREX**, for tRNA Extension.

The validation experiment proved that aminoacylation of $tRNA^{Pyl}$ generated a species which was protected from $NaIO_4$ oxidation when assayed by tREX. This result led us to ask how the relative intensity of the two bands resolved on the gel were modulated as a function of the aminoacylation level of $tRNA^{Pyl}$. Urea PAGE analysis had revealed that even at high concentrations (4 mM) of N^{ε}-Boc-L-lysine, aminoacylation of $tRNA^{Pyl}$ by PylRS is not complete (**Figure 2.6a**). However, the control experiments used to validate tREX suggested that a tRNA sample which is not oxidised before tREX behaves as a completely aminoacylated sample, thus being fully extended (**Figure 2.5** lanes 10). I hence designed an experiment in which I used oxidised or unoxidised *E. coli* tRNA extracts containing $tRNA^{Pyl}$ as an equivalent for 0% and 100% aminoacyl-$tRNA^{Pyl}$ over $tRNA^{Pyl}$, respectively. I then mixed these samples in known ratios to obtain apparent aminoacylation levels corresponding to 0%, 5%, 15%, 25%, 50%, 75% and 100%, and used these as a calibration curve (**Figure 2.6c**) to estimate the fraction of BocK-$tRNA^{Pyl}$ present in tRNA extracts from cells grown in presence of 1 mM, 2 mM or 4 mM of BocK (**Figure 2.6b**).

The data showed that the relative fluorescence intensity from the upper and lower bands correlates with the relative fraction of aminoacyl-$tRNA^{Pyl}$ equivalents present in the sample and that the same trend is observed when the cell extracts were tested by tREX contained increasing levels of BocK-$tRNA^{Pyl}$. Comparison of the biological samples to the calibration curve generated as described allowed me to estimate the fraction of aminoacylation as a function of the concentration of amino acid in the medium, and these results correlated well with the estimate made by quantification of the

norther blot performed after urea PAGE (**Figure 2.6d**). The difference in the absolute numbers obtained by the quantifications using different methods was probably due to experimental error on the one hand, and on the fact that while tREX measurement are compared to a calibration curve, quantification on the northern blot takes into account the nominal value of pixel intensity due to the unavailability of a sample containing only BocK-tRNA^{Pyl} required to generate a calibration curve using this method. Overall, these results indicate that tREX can be used quantitatively to measure the aminoacylation fraction of a specific tRNA, even if for the project the assay was used mostly qualitatively as described in the next section.

Overall, tREX is independent of the anticodon of the tRNA under analysis by design, is independent of the nature of the amino acid bound to the tRNA of interest since any amino acid confer equal resistance to $NaIO_4$ oxidation, and is easily scalable to allow investigation of multiple tRNA in parallel

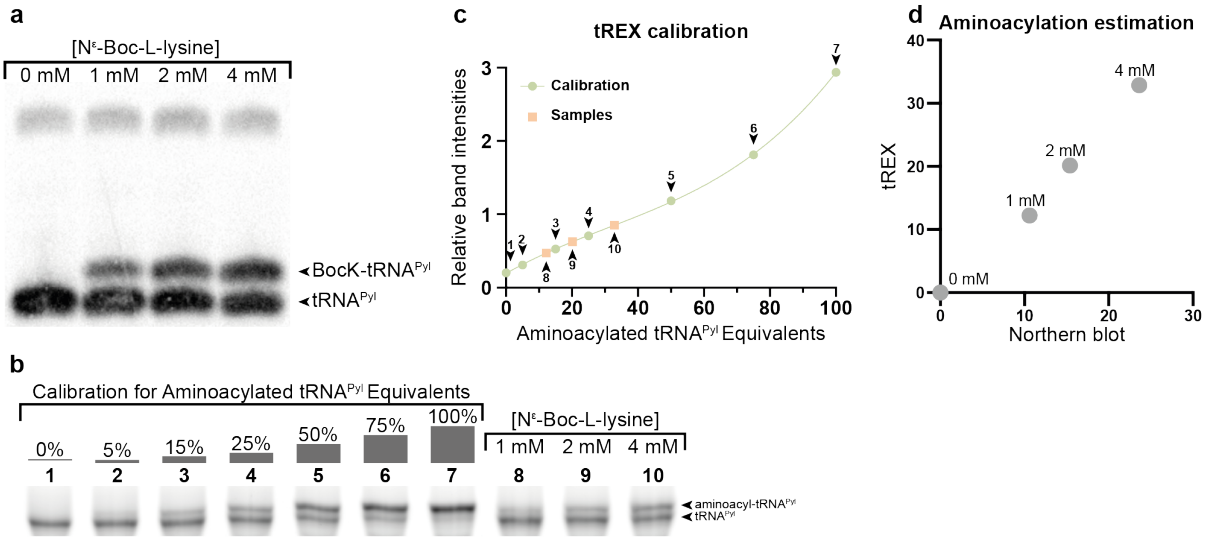


***Figure 2.6**: Amino acid-dependent fractional tRNA aminoacylation measured by northern blotting or by tREX. **a)** Growth of E. coli cells expressing tRNA^{Pyl} and PylRS in presence of increasing concentrations of N$^{\varepsilon}$-boc-L-lysine in the range from 0 to 4 mM of amino acid led to an increase of the fraction of aminoacylated tRNA^{Pyl} which can be observed by urea PAGE followed by northern blot. **b)** The relative intensity of the extended vs. non-extended bands produced by tREX correlated with the level of aminoacylation of the tRNA under investigation. Mixing of a tRNA extract oxidised following complete deacylation (0% aminoacylation) together with a tRNA extract for which the $NaIO_4$ oxidation was omitted (100% aminoacylation) allows the generation of a calibration curve (third degree polynomial) used as a reference to estimate the fractional aminoacylation of tRNA extracts from cells grown in presence of variable concentrations of N$^{\varepsilon}$-boc-L-lysine. **c)** Calibration curve derived from the data shown in b) to estimate the fractional aminoacylation of tRNA^{Pyl} as a function of the concentration of BocK. Lanes 1-7 in b) (green circles) were used to generate a curve. Lanes 8-10 in b) (orange squares) were fitted on the curve to estimate the fractional aminoacylation of the samples. **d)** Estimation of the fractional level of aminoacylation of tRNA^{Pyl} in extracts from cells grown in presence of variable concentrations of N$^{\varepsilon}$-boc-L-lysine from the northern blot in a) and from the tREX data in b) showed a high level of correlation. The difference in the absolute values of aminoacylation fractions measured by the two methods might be due to the impossibility of generating a calibration curve for northern blots, as this would require the availability of a tRNA sample fully aminoacylated. Conversely, tREX allows for the generation of a calibration curve which can be used as a standard for quantification.*

as tRNAs only need to be oxidised before the procedure, while the probe hybridisation and tRNA extension reactions can be performed serially in one pot with no requirement for purification or additional processing before electrophoresis. Furthermore, the detection is performed immediately following electrophoresis by *in gel* fluorescence. Consequently, we reasoned that tREX could be a suitable method to perform the screening to identify tRNAs which, when introduced in *E. coli*, are expressed but not aminoacylated by any of the *Ec*-aaRSs.

Screening for tRNA Orthogonality

Having developed an effective assay to test aminoacylation of a particular tRNA, we could proceed to address the next goal of the project, namely identify new tRNAs which are expressed and orthogonal in *E. coli* (**Figure 1.3 (2)**). However, tREX had only been successfully used to measure the aminoacylation of a single tRNA, hence we decided to first apply the new method to test a subset of 10 candidate tRNAs (**Figure 2.7** Trp_11 to Trp_20). We purchased each of the candidate tRNA genes, including the 5'-CCA-3' universal sequence at positions 74→76, cloned into the same position of a standard plasmid under the control of the *E. coli lpp* promoter, which drives high transcription of its downstream genes. For each candidate, a Cy-5 labelled DNA probe was designed to be complementary to the portion of its target between positions 45 and 76 and the length of the poly-A tail at its 5'-end of the probe was adjusted based on the length of the tRNA, in order to obtain a visible electrophoretic retardation in every case (**Chapter IV – Materials & Methods: tREX Probes Design**). Since every sample is tested with a different probe, it was necessary to verify that the signal observed in each case was specific. For this reason, a wild type *E. coli* tRNA extract was oxidised and used as a specificity control for every probe. Also, for each different tRNA tested it was important to generate controls for the electrophoretic mobility of the tRNA/probe heteroduplex and for the fully extended species. In order to do this. As discussed before, these two controls could be generated by either oxidising a fully deacylated tRNA sample or by omitting the oxidation step altogether, respectively. Differently from the case of PylRS/$tRNA^{Pyl}$, aminoacylation for the $tRNA^{Trp}$ under investigation cannot be modulated by the addition of an amino acid in the medium. For this purpose, chemical deacylation was induced by incubating the tRNA in an alkaline environment due to the liability of the aminoacyl ester bond in basic pH. Consequently, for each candidate under analysis, tREX was performed on 4 samples:

(i) the first is the wild type *E. coli* tRNA extract;

(ii) the second was the tRNA extract on which oxidation was omitted (i.e. control for the mobility of the fully extended species);

(iii) the third was the tRNA extract which oxidised following chemical deacylation (i.e. control for mobility of the unextended species);

(iv) the fourth was the oxidised extract.

In order to identify an orthogonal tRNA, the sample (*iv*) must be indistinguishable from the control (*ii*) and must lack the extended band present in sample (*iii*). In addition, for tRNAs which are expressed in *E. coli* and which can be detected specifically by tREX, the signal observed in samples (*ii*), (*iii*) and (*iv*) must be absent from the control (*i*).

My first test confirmed that specific detection could be observed for all $tRNA^{Trp}$_11 to $tRNA^{Trp}$_20 with the exception of $tRNA^{Trp}$_18 (**Figure 2.7**). In this case, the signal in all the samples (*i*) to (*iv*) was indistinguishable and I concluded that this tRNA was not detectable. Considering the high binding energy between complementary DNA and RNA sequences, the absence of specific signal was more likely due to the lack of expression of the tRNA, even if lack of binding could not be excluded. In the other cases, specific bands of various degrees of fluorescence intensities could be observed in the gel, indicating aminoacylation of all the samples to very high levels (e.g.: $tRNA^{Trp}$_12). In several cases, a pattern of bands was observed rather than a single band corresponding to the tRNA of interest. These bands could correspond to maturation intermediates of the tRNA, or degradation products which might be generated either *in vivo* or *in vitro* during the tREX protocol. However, since every sample is analysed from the comparison to a set of controls generated following the same procedure, the presence of these patterns does not affect the interpretation of the results.

Having verified that tREX had broader applicability on various heterologous tRNAs, we proceeded by expanding the screening to 233 other candidate tRNAs in addition to the 10 test $tRNA^{Trp}$ above mentioned. In order to speed up the process, a part of the screening was performed by Julian. C. W. Willis and Louise H. Funke. A subset of the screening gels are exemplified here (**Figure 2.7**). The complete set of gels generated is also reported (**Figure 5.1**). A summary of the result of the screening is shown in **Table 2.1**.

Overall, based on the experimental data we concluded that 75 out of 243 tRNAs (~30.9%) could not be detected in the tRNA extract. In these cases, no conclusions could be drawn on whether such tRNAs would be effective substrate for *E. coli* aaRSs, hence these tRNAs provide little insight about how to refine our procedure. Of the remaining 168, 97 (~57.7%) displayed a detectable amount of aminoacylation, while in 71 cases (~42.3%) the oxidised tRNA extract was showed a fluorescence pattern indistinguishable from the chemically deacylated control (*iii*) and was for this reason considered orthogonal.

	Isoacceptor class										
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Total**
Undetectable	6	2	5	1	3	2	1	2	3	4	29
Aminoacylated	3	4	3	-	1	5	3	4	1	3	27
Orthogonal	6	4	2	8	4	3	6	4	6	3	46
Total	15	10	10	9	8	10	10	10	10	10	102

	Isoacceptor class										
	Leu	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**	**Total**
Undetectable	14	3	1	1	3	10	3	5	3	3	46
Aminoacylated	14	6	9	7	3	3	2	13	6	7	70
Orthogonal	-	-	-	2	3	12	5	2	1	-	25
Total	28	9	10	10	9	25	10	20	10	10	141

***Table 2.1**: Summary of the screening for tRNA orthogonality performed by tREX.*

Interestingly, the distribution of orthogonal tRNAs among different isoacceptor class was

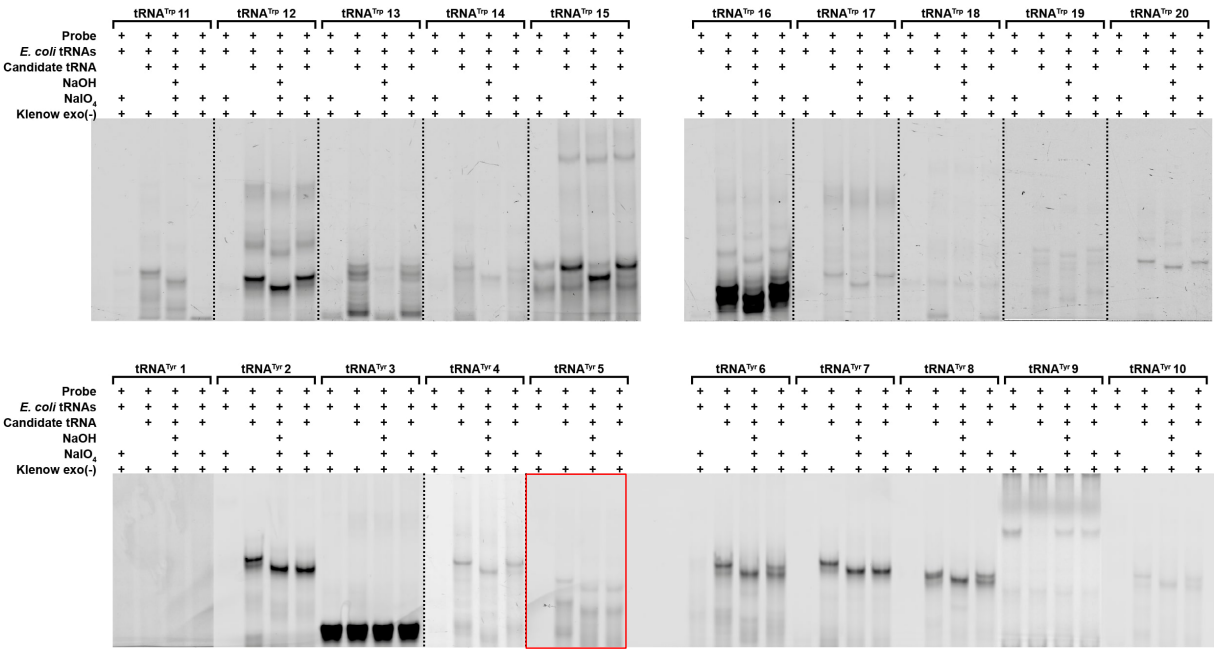


***Figure 2.7**: Example gels containing the results of the tREX screening for tRNA orthogonality. For each tRNA, a specific DNA probe was designed to be complementary to it. Furthermore, the specificity of each probe was tested on a tRNA extract from wild type E. coli DH10b. For each tRNA under investigation, a control for the electrophoretic mobility of the extended species was generated by omitting the $NaIO_4$ oxidation, while a control for the electrophoretic mobility of the unextended species was generated by performing $NaIO_4$ oxidation following chemical deacylation by NaOH. Orthogonality was assessed by verifying if the oxidised tRNA sample displayed the a band with the same electrophoretic mobility as the control for the unextended species and no bands with the same electrophoretic mobility as the control for the extended species. In the example shown, only the $tRNA^{Tyr}5$ was detectable but not aminoacylated in E. coli and is highlighted by a red box.*

significantly uneven. As an example, orthogonal $tRNA^{Asp}$ were found in very high abundance among the candidates tested (8 out of 9), while only one candidate $tRNA^{Tyr}$ was orthogonal. For four isoacceptor classes (i.e.: Leu, Lys, Met and Val) none of the candidates which could be detected were orthogonal. In spite of this, the 71 orthogonal tRNA identified were distributed among 16 distinct isoacceptor classes, indicating that tRNAs from species on a different evolutionary line compared to the *E. coli* might have developed distinct recognition mechanism by their cognate aaRS while lacking the interaction elements required for *Ec*-aaRSs.

This screening represented the first type of systematic analysis of tRNA orthogonality as far as we were aware of. Importantly, the high frequency at which orthogonal tRNAs could be identified provided us with a large number of candidates, which we used to move to the third step of the project, namely identifying aaRSs which could effectively aminoacylate these new orthogonal tRNAs in *E. coli* (**Figure 1.3 (3)**).

Identifying Active aaRS/tRNA Pairs

For each of the tRNAs which were expressed in *E. coli* but could not be aminoacylated by any of its aaRSs, we asked whether the cognate synthetase from the same organism where the tRNA was identified could aminoacylate this tRNA in *E. coli*. tRNA-DB-CE contains information about genome ID on which the tRNAs were informatically identified, together with the name of the species to which the genome belongs to. However, I found that in many instances this species annotation was not up-to-date, which made the identification of the aaRSs troublesome. In order to ensure the highest accuracy in the identification of the enzymes to use in combination to each tRNA, I decided not to rely on the taxonomic information from the tRNA-DB-CE but to retrieve such information from the NCBI databases.

I searched the genome ID on NCBI Genomes to retrieve taxonomy ID of the corresponding species from NCBI Taxonomy. This identifier allowed me to retrieve a complete and updated set of taxonomical information about the species, and importantly it also allowed me to unambiguously identify a given organism even in the circumstance in which its phylogenetic and taxonomic affiliation were modified, which is not unfrequent for newly discovered organisms. Subsequently, I searched NCBI Proteins to identify the sequence of the enzyme annotated as the cognate aminoacyl-tRNA synthetase in that organism. In all cases, I could find a protein annotated with the required enzymatic activity and I did not have to perform homology searches on unannotated or predicted

proteins.

I reverse translated the protein sequences using the canonical genetic code to obtain CDSs codon-optimised for *E. coli* using the IDT codon optimisation tool (https://www.idtdna.com/CodonOpt). The sequences were designed taking into account the additional constraint that they should not contain BsaI restriction site, commonly used to perform enzymatic reverse PCR to create library. The linear dsDNA for each synthetase was purchased from Twist Bioscience and cloned into the plasmid containing its respective tRNA gene. No tags were fused in frame with the CDS for the aaRSs, to prevent the possibility that alteration in the protein sequence could result in a loss of activity of the enzyme. The resulting plasmids were transformed in *E. coli* and the tRNA extracts from these cells were tested again using tREX. In this case, for each tRNA under investigation 5 samples were tested using tREX:

i) the wild type *E. coli* tRNA extract;

ii) the tRNA extract on which oxidation was omitted (i.e. control for the mobility of the fully extended species);

iii) the tRNA extract oxidised following chemical deacylation (i.e. control for mobility of the unextended species);

iv) the oxidised tRNA extract from cells expressing both the candidate tRNA of interest and its cognate aaRS;

v) the oxidised tRNA extract from cells expressing the candidate tRNA alone.

The result of the screening are summarised in **Table 2.2** and shown in **Figure 5.2**, while a subset of gels is shown in **Figure 2.8**. Among all the pairs tested, the gels showed unambiguous aminoacylation of 23 out of 59 orthogonal tRNAs when their cognate aaRS was present in the cell, but no aminoacylation in its absence, reproducing the data shown before. Consequently, I was able to identify 23 active aaRS/tRNA pairs which are active when heterologously expressed in *E. coli* and for which the tRNA cannot be recognised by any of the cellular *Ec*-aaRSs. Interestingly, for some isoacceptor classes (e.g.: Asp, Cys, Gln, Glu, Ile) all of the tested pairs showed clear activity, while for others (e.g.: Ala, His, Ser) the success rate was very low or null. It is relevant to notice that the lack of aminoacylation of the orthogonal tRNAs by their cognate synthetase cannot specifically be pinned down to a particular cause, as several factors might be responsible for the lack of activity of the aaRSs. For example, the enzymes might be expressed poorly, might not fold correctly, might be inactive at the particular physiological condition present in the *E. coli* cytosol (e.g.: temperature, salt concentrations etc.), or their cognate tRNAs might be poorly folded or incorrectly modified.

Overall, the 23 active pairs were distributed among 10 different isoacceptor classes, thus covering

half of the natural proteinogenic amino acids. This experiment further highlights the value of tREX as a method to test *in vitro* the aminoacylation which happens *in vivo*.

Having assessed the activity of these 23 synthetases brought us closer to the identification of new orthogonal pairs. However, the experiment shown so far were not enough to assess whether the newly tested enzymes were able to effectively aminoacylate the endogenous *Ec*-tRNAs. As discussed previously, it has been impossible to date to design an experiment which could assess orthogonality for exogenous aminoacyl-tRNA synthetases for canonical amino acids *in vivo* under physiological conditions. For this reason, I performed *in vitro* aminoacylation assays, as described in the next section.

	Isoacceptor class										
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Total**
Inactive pair	6	2	2	-	-	-	-	2	5	-	17
Active pair	-	2	-	3	4	3	5	1	1	2	21
Total	6	4	2	3	4	3	5	3	6	2	38

	Isoacceptor class										
	Leu	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**	**Total**
Inactive pair	-	-	-	2	2	10	4	1	-	-	19
Active pair	-	-	-	-	1	-	-	-	1	-	2
Total	-	-	-	2	3	10	4	1	1	-	21

***Table 2.2**: Summary of the results shown in **Figure 2.8**.*

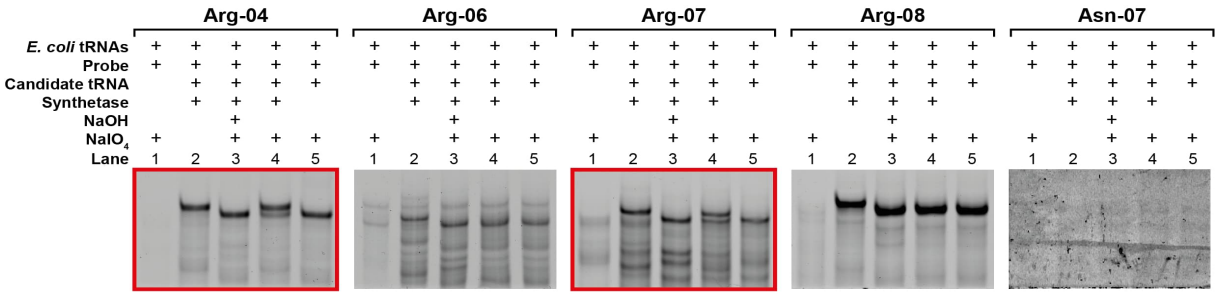


***Figure 2.8**: Example set of gels containing the results of the tREX screening for the aminoacyltion of the orthogonal tRNA by their cognate synthetase. For each tRNA under investigation, a control for the electrophoretic mobility of the extended and unextended species was generated as described before. The aminoacylation status of the tRNA was verified when it was co-expressed in E. coli cells together with its cognate synthetase from the same organism. Samples for which a synthetase-dependent aminoacylation was observed are highlighted by a red box.*

Testing aaRS Orthogonality

In order to successfully expand the genetic code while ensuring a high fidelity for protein translation, the aminoacyl-tRNA synthetases present in the cell together must all be mutually orthogonal, i.e. they should have no overlap in the set of tRNAs which they can effectively aminoacylate. As outlined in **Figure 1.3**, having successfully completed the tasks **(1)**, **(2)** and **(3)**, the last step required to verify that the approach described was successful in identifying new orthogonal pair was the assessment of orthogonality of the 23 synthetases which were proven active on their cognate orthogonal tRNAs. When testing aaRS orthogonality *in vitro*, this effectively meant that what needed to be verified was that the new synthetases could aminoacylate a tRNA extract from *E. coli* exclusively if the extract contained their cognate tRNA.

To set up this experiment, I cloned 15 of the active aaRS in a commercial pET expression vector, where the protein of interest is under the control of the strong T7 promoter and could be expressed in *E. coli* BL21 strain upon induction of the T7 RNA polymerase with IPTG. The proteins were N-terminally fused in frame with a StrepTag II sequence followed by the cleavage site for the TEV protease (**Chapter IV – Materials & Methods: Synthetase Purification**). This design enabled one-step purification of all the enzymes to high purities, while cleavage of the tag following purification was performed to ensure that the activity of the enzyme was not compromised by the addition of the tag. In addition, the *M. jannaschii* TyrRS was purified similarly and used as a positive control.

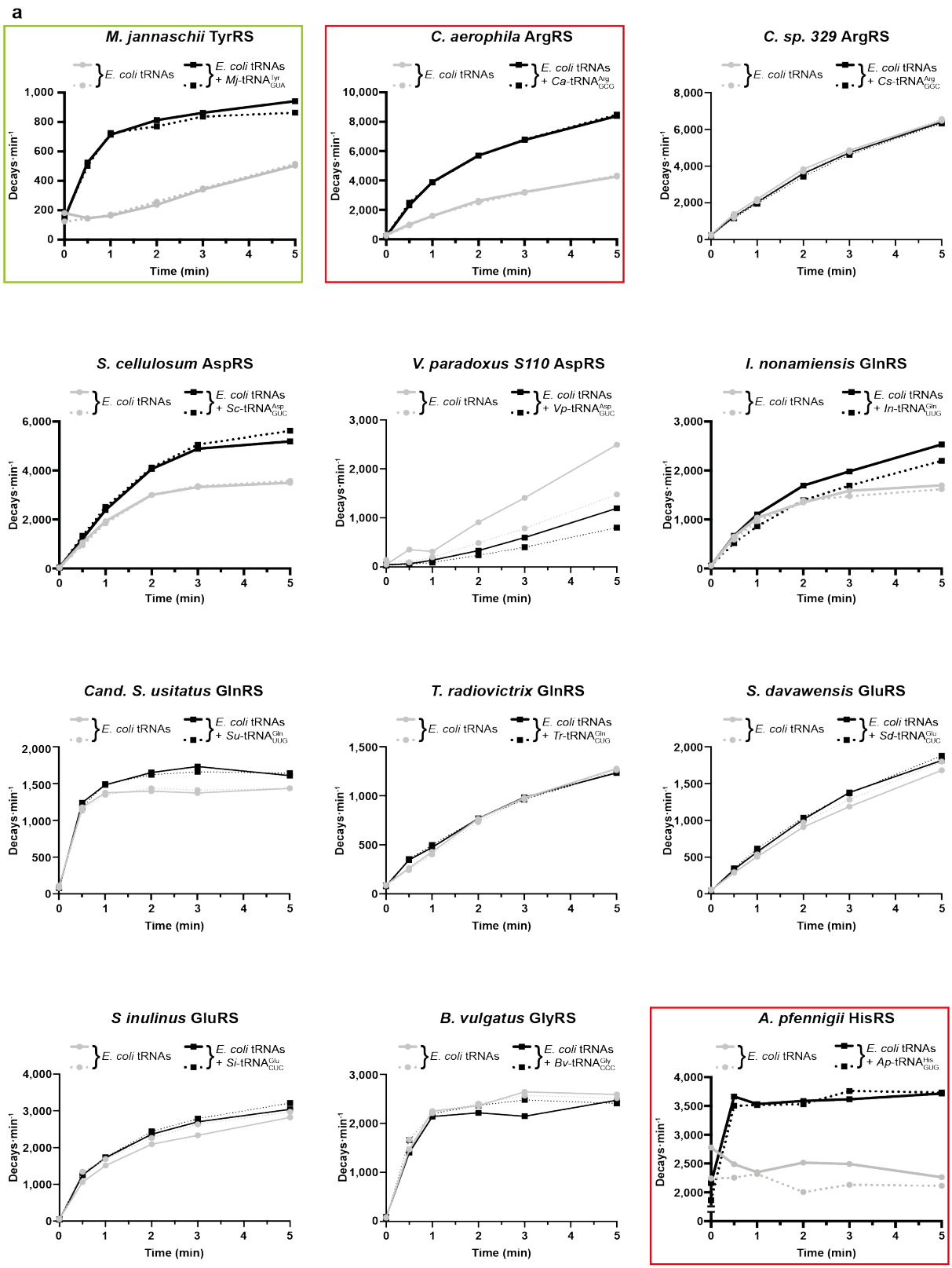
To obtain tRNAs of the highest quality possible, I decided not to produce them by *in vitro* transcription, but to purify them from *E. coli* instead. This had the advantage of ensuring that all the tRNAs were correctly folded, modified and present at the physiological relative proportions. To do this, I transformed *E. coli* DH10b with the plasmids containing each tRNA/aaRS pair which I used to perform the experiment in **Figure 2.8**, then I extracted the tRNAs, chemically deacylated them by treatment with NaOH and then performed extensive desalting using size exclusion spin concentrator columns (**Chapter IV – Materials & Methods: tRNA Extraction for in vitro Biochemistry**). These extracts were effectively used as positive controls, because the activity of the synthetases for their cognate tRNA was verified *in vivo*, and for this reason I expected a positive signal from the *in vitro* assay. Additionally, tRNAs from wild type *E. coli* DH10b were purified in the same way and were used to verify whether the aaRS under investigation could charge any of the endogenous tRNAs or not. As for orthogonal tRNAs no signal was expected in this assay, the presence of the positive control

ensured that the lack of signal wasn't due to the inactivity of the enzyme *in vitro*.

In my first attempts to test orthogonality, I assumed that for the well characterized orthogonal *Mj*-TyrRS no aminoacylation on the wild type *E. coli* extract would occur after incubation with L-tyrosine and ATP, while the reaction should occur when the tRNA extract contains the *Mj*-$tRNA^{Tyr}$. Since aaRSs produce AMP and PP_i as a by-product of their enzymatic activity, I originally tried to measure the production of AMP by the commercial kit AMP-Glo™ (Promega) produced in the reaction after 1 h incubation at 37°C. Unfortunately, I observed a strong signal above background from both sample (data not shown). This indicated that while *Mj*-TyrRS are specific for its substrates *in vivo*, the basal affinity for tRNAs is high enough so that, following a long enough incubation period, aminoacylation of non-specific targets occurs. As a consequence, I evaluated end-point measurements as ineffective and opted to perform a more traditional assay which allowed to measure time courses over a short time window. Importantly, when maintaining the concentrations of the substrates equal, the relative initial reaction rate in a time courses correlates with the affinity of the enzyme for its substrate, as higher affinity results in higher initial reaction rate.

The assay employed to measure aminoacylation takes advantage of the difference in solubility in acidic pH between amino acid and nucleic acids. In fact, while amino acids remain in solution upon addition of strong acids (e.g.: trichloroacetic acid), tRNAs quickly precipitate out of solution. Importantly, as the aminoacyl ester bond is stable at low pH, acid can be used to selectively precipitate the amino acids bound to tRNAs, hence quantification of the amount of amino acid which co-precipitates with the tRNAs at different time points of the aminoacylation reaction is a dynamic measurement of the activity of the enzyme.. Radioactive amino acids were purchased to allow direct measurement by scintillation counting. ^{13}C-radiolabelled amino acids (Arg, Asp, Gln, Glu, Gly, Ile, Pro, Tyr) were purchased, while histidine was purchased ^{3}H-radiolabelled due to supply availability. Unfortunately Cys was not available from the supplier. The reactions were set up by incubating the appropriate tRNA extracts with each of the purified synthetases and their corresponding labelled amino acid, then started by adding ATP. Aliquots of the reactions at different time points were precipitated on glass filters soaked with TCA and the radioactivity precipitated in the filters was measured following removal of the unbound amino acids. The reactions were monitored over the course of 5 minutes and the results are shown (**Figure 2.9a**). For each enzyme, two time courses were performed on tRNA extract lacking the cognate tRNA (grey traces) and two other time courses were performed in its presence (black traces) as positive controls. All the samples preparations displayed high purify following a single step of affinity purification (**Figure 2.9b**) and were enzymatically active. These experiments were performed with the assistance of Dr. Shan Tang.

Surprisingly, among all the synthetases tested, only 3 out of 15 showed an appreciable difference in aminoacylation kinetics between the two conditions (**Figure 2.9a**, red boxes), similarly to what



observed for the positive control (**Figure 2.9a**, green box): the ArgRS from *Caldilinea aerophila*, the HisRS from *Afifella pfennigii* and the TyrRS from *Archaeoglobus fulgidus*.

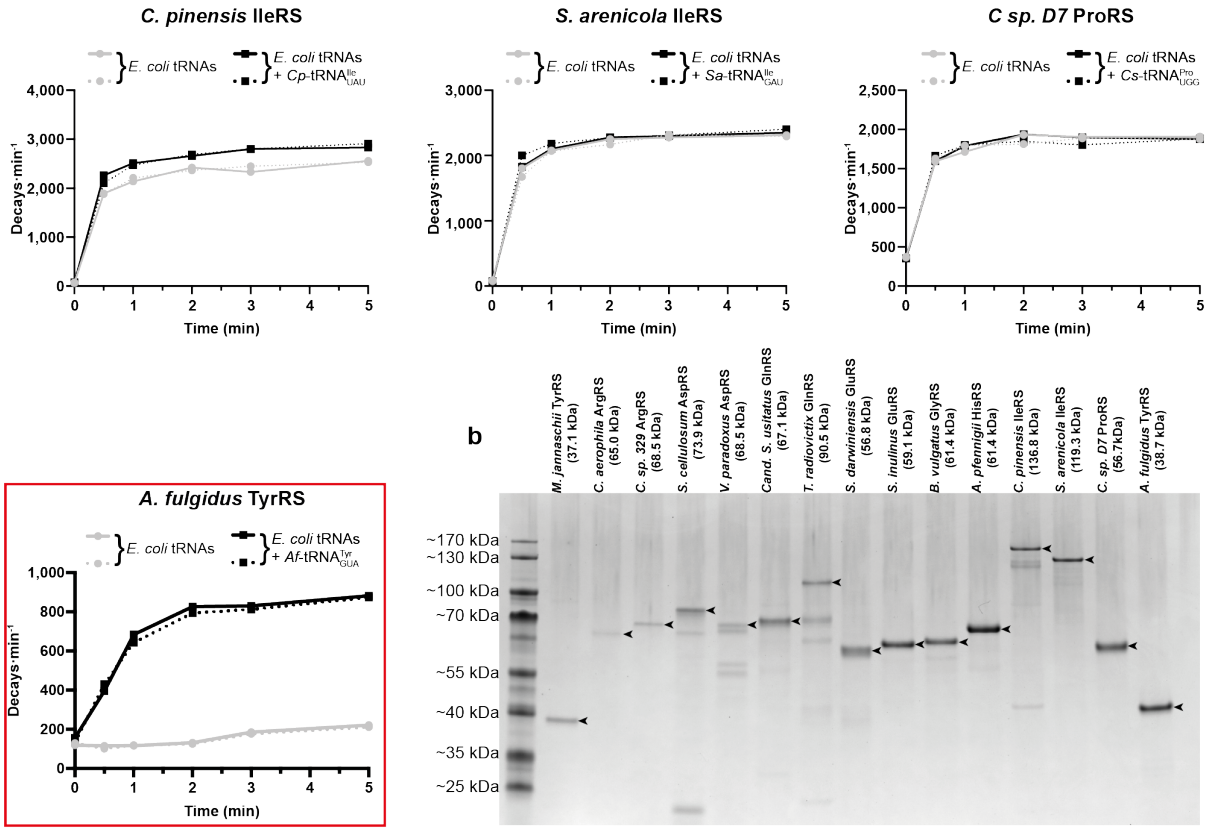


***Figure 2.9**: **a)** In vitro aminoacylation assays. Each aaRS was purified and its activity was assayed by measuring the time course of the incorporation of its radioactively-labelled substrate amino acid into the acid-insoluble tRNA fraction. The synthetases were incubated with either a tRNA extract from wild type E. coli (gray traces), or with a tRNA extract from E. coli expressing its cognate tRNA (black traces) as a positive control. Orthogonality of the synthetases by this method was assayed by the difference in the kinetics of the incorporation of amino acid in the presence or absence of a specific substrate for the enzyme. Enzymes which displayed a negligible or slow aminoacylation of the wild type E. coli tRNA extract compared to the tRNA extract containing their specific substrate were considered orthogonal (red boxes). The aminoacylation kinetic for the known orthogonal Mj-TyrRS was measured as a positive control (green box). **b)** SDS-PAGE of the purified enzymes used in the in vitro assays revealed high levels of purify following a single affinity purification step. The band corresponding to the full-length protein is indicated by an arrow. I. Nonamiensis GlnRS could not be included as all the material was used in the in vitro assays.*

This result was particularly surprising to me, in light of a few considerations. Firstly, all the enzymes tested could be successfully expressed and purified from *E. coli* and the cell from these cultures did not display any signs of a detrimental activity of these exogenous proteins. This highlighted how these synthetases could not be recognising to a significant extent any of the *E. coli* tRNAs for a distinct isoacceptor class, as this would result in a loss of fidelity in protein translation, hence a loss of fitness for those cells. As pointed out before, though, recognition of the *E. coli* tRNA for the same isoacceptor class would not induce a loss in viability of the cell, thus being the most likely event for a non-orthogonal and non-toxic synthetases. Another important consideration to make concerned the mechanism of recognition between the exogenous aaRS/tRNA pair compared to the *E. coli* ones.

Let us take into account the ArgRS/tRNA^{Arg} pair from *Capnocytophaga sp. oral taxon 329*, a species belonging to the phylum *Bacteroidetes*, hence evolutionarily distant from *Proteobacteria*, the phylum in which *E. coli* is found. This tRNA was clearly detectable in *E. coli* but displays no sign of aminoacylation (Arg-04 in **Figure 2.8**). This indicated that, since the last common ancestor, the *Cp*-tRNA^{Arg} and the *Ec*-tRNA^{Arg} had evolved and diverged in such a way that the features recognised by *Ec*-ArgRS are present in the *Ec*-tRNA^{Arg} but not in *Cp*-tRNA^{Arg}. However, **Figure 2.8** clearly shows that *Cp*-ArgRS can aminoacylate its cognate *Cp*-tRNA^{Arg} in *E. coli*. Since the features recognised by *Ec*-ArgRS are absent from *Cp*-tRNA^{Arg}, then *Cp*-tRNA^{Arg} must display a different set of features which *Cp*-ArgRS, but not *Ec*-ArgRS, can recognise. In other words, *Cp*-ArgRS must have evolved a new recognition method to interact with its partner *Cp*-tRNA^{Arg} which relies on a new set of features, different from the ones which *Ec*-ArgRS recognises on the *Ec*-tRNA^{Arg}. Under these circumstances, and considering that *Ec*-ArgRS cannot recognise the *Cp*-tRNA^{Arg}, I considered unlikely that:

a) *Cp*-tRNA^{Arg} only displays the features recognised by *Cp*-ArgRS;

b) *Ec*-tRNA^{Arg} displays both the features recognised by *Ec*-ArgRS and by *Cp*-ArgRS.

In fact, since these pairs are found in evolutionarily distant species, this would imply the existence of an unusual asymmetry in the evolutionary path undertaken by the two different tRNA sequences. For this reason, I imagined that the probability of identifying orthogonal synthetase from a selected pool of pairs in which the tRNA was known to be orthogonal in *E. coli* should be significantly higher than the one observed. In addition, for most synthetases there was no apparent difference in the aminoacylation kinetics between the sample containing their cognate tRNA and the wild type *E. coli* tRNA extract. This would suggest not only a rather counterintuitive lack of differential affinity for its substrate rather than other tRNAs, but also that the total amount of substrates in the two sample was equal, which was also counterintuitive due to the additional presence of a specific substrate in one of the two samples.

Overall, the results obtained from these experiments were not completely satisfactory and posed some questions on the possibility of deducing *in vivo* orthogonality from *in vitro* assays. However, due to the lack of possible alternatives, we could not investigate the *in vivo* orthogonality of the aaRSs by other means.

Discussion

In the previous sections I have described how we decided to approach the problem of establishing a general procedure which could allow us to identify tRNA/synthetase pairs derived from known living organisms which are orthogonal when heterologously expressed in *Escherichia coli*. First, we decided to take advantage of an annotated database of bacterial and archaeal tRNA sequences, tRNA-DB-CE, which contained more than 2 million sequences in March 2017. Our choice to limit our search by excluding pairs derived from eukaryotes was mostly due to the practical need to verify the reliability of our newly developed pipeline on an initial set of tRNAs which wouldn't be excessively large, and to the consideration that eukaryotic synthetases can be clustered in multi-enzymatic complexes[117], which would introduce additional complications along the pipeline. Additionally, we did not include tRNA annotated on fragmented and only partially assembled genomes derived from metagenomics on environmental samples due to the possibility of not being able to retrieve the corresponding synthetase for these tRNAs. Lastly, as the tRNA genes present in the database were derived from computational predictions, we made use of the genes which were annotated as "reliable tRNA genes" by the database curators, which correspond to genes which were predicted to code for tRNAs by multiple distinct computational tools.

Having selected an initial set of sequences, we decided to develop a computational filtering method which would reduce the complexity of our tRNA set by identifying the ones with the highest likelihood of being orthogonal in the host we chose for our experiment. In order to do this, we took advantage of the knowledge of how *E. coli* synthetases interact with their partner tRNAs. In particular, in a first approximation it is believed that each of the 20 distinct cellular aminoacyl-tRNA synthetases recognise and establish specific interactions with a defined sub-set of the nucleotides along the sequences of the cellular tRNAs, known as identity elements[20]. If the sequence of a given tRNA at the identity elements for a given synthetase is incorrect, the model predicts the interaction to be negated. This model clearly does not take into account the contributions to the binding energy between a tRNA and a synthetase given by complex three-dimensional features, like the relative orientation of different portions of the tRNA and the length of some of its variable loops. However, the simplicity of this model allowed us to generate a scoring system which, even if not predictive of the interactions between tRNAs and the *E. coli* synthetases with 100% accuracy, could increase our chances of identifying an orthogonal tRNA compared to the chance of finding orthogonality in a

completely random selection of tRNA genes.

Our filter compared the sequence of any given tRNA at the identity element for a specified *E. coli* aaRS with the sequence at the same positions of a known substrate for that aaRS, identified in the cognate *E. coli* tRNA(s), assigning a score of +1 to all those identity elements where the tRNA matched the synthetase's substrate, and a score of -1 to the ones which were different. We calculated the overall score of a given tRNA for a specified *E. coli* aaRS as the average of its scores across all the identity elements for that synthetase. Importantly, this procedure was possible thanks to the literature which identified the identity elements for the synthetases present in the host organism of choice, i.e. *E. coli*[20, 118-120]. Since the enzymes present in different organisms differ in their method of substrate recognition, the list of identity elements used by the filter would have to be adapted if the procedure was to be repeated to identify tRNAs which are orthogonal in another organism.

Given its design, the scoring system marked with scores close to +1 tRNAs which were very similar to the endogenous substrate for the *E. coli* synthetases, while tRNAs with scores close to -1 were the most dissimilar. We could hence identify a sub-set of the original database which was enriched in tRNA sequences which we considered more likely to be orthogonal.

To obtain experimental evidence of the orthogonality of tRNAs, I developed a new method which would address some of the important limitations which the traditional methods display when used to perform screening. In particular, ease of execution and scalability represented a critical concern for the application of acidic urea PAGE followed by northern blotting, which represents to date the golden standard to verify tRNA aminoacylation. By combining the specificity of $NaIO_4$ oxidation on vicinal diols with RNA detection by fluorescently labelled DNA probes, I developed a new method, which I called tREX, which is able to induce a reduction in the electrophoretic mobility of aminoacylated tRNA, resolving them from their free counterparts. I successfully validated the method, proving that when measuring aminoacylation of $tRNA^{Pyl}$, two bands were resolved on a polyacrylamide gel whose relative intensity depended on the fractional level of aminoacylation of the tRNA. Importantly, tREX can be performed in PCR tubes, allowing for convenient scalability. Even if significantly easier to perform on a few hundreds of samples than acidic urea PAGE/northern blot, tREX still required over a thousand lanes of gels to be run, and even if several samples were screened in parallel, the technique requires a good amount of labour. Given its robustness, though, future improvements in the method, like automation of some of its steps, might increase its throughput.

By using tREX, 243 different tRNAs were tested for their aminoacylation status when heterologously expressed in *E. coli*. Of those, 168 could be specifically detected: 97 displayed some degree of aminoacylation, while for 71 no detectable aminoacylation could be observed (~42%) and were considered orthogonal. tREX was also used to determined whether co-expression in *E. coli* of one of those orthogonal tRNAs with the cognate synthetase from the same organism of origin could result in

aminoacylation of the tRNA itself, and indeed we observed that this was the case for 23 out of the 59 pairs (~39%) we tested. To our knowledge, this was the first large-scale attempt to characterise orthogonality in *E. coli* of tRNAs derived from natural species preserving their wild type anticodon. Surprisingly, a substantial fraction of all the tRNAs which could be expressed in *E. coli*, belonging to 16 out of the 20 isoacceptor classes, were not aminoacylated by any of its endogenous synthetases. We hypothesised that this success rate was higher than what we would have observed by randomly selecting a tRNA to express in our host organism. This would indicate that, in spite of its simplicity, and of its partial predictive power, our scoring method can be effectively used to filter tRNA genes and enrich for orthogonal sequences. Nonetheless, it is likely that a fraction of those sequences which were filtered out by our algorithm are orthogonal in *E. coli*. The experimental data generated by tREX provided us with an expanded set of sequences which are validated to be orthogonal. This information will probably be used in the future to polish our computational approach and to widen its scope to a wider set of tRNAs, taking advantage both of the progress made in the sequencing and assembly of bacterial and archaeal genomes, and also of databases of eukaryotic tRNA genes. While the sequences of tRNAs which were not orthogonal in *E. coli* could represent an important information towards the refinement of the algorithm as well, no experimental data was collected about the amino acid with which they were charged, so that we cannot unambiguously identify the endogenous synthetases by which they were recognised. Overall, our data proved that tRNA orthogonality is not a rare occurrence and that the natural divergence generated a very broad plethora of different genes with distinct features, enabling the evolution of distinct modality of recognition by their cognate synthetases. Considered the impressive variety of sequences available to date, it is easy to imagine how multiple orthogonal tRNAs must exist for every one of the 20 natural isoacceptor classes.

The identification of 23 pairs composed of an orthogonal tRNA and an active synthetase implies that close to 9.5% of all the tested tRNAs, or approximately 13.7% of all the expressed tRNA, ended up constituting one such pair. The existence of orthogonal tRNAs which could not be aminoacylated by their cognate synthetases suggested that, while being effectively transcribed, some of those tRNAs might have been incorrectly folded or modified, preventing their interaction with their partner enzyme. Alternatively, the expression, folding or multimerisation of the synthetases in *E. coli* could be defective.

Given the problems in proving orthogonality for synthetases *in vivo*, we opted for a less conclusive *in vitro* approach which had been used already before[40, 93, 94, 97]. To my surprise, the assay seemed to indicate that only three of the tested synthetases were orthogonal. The large amount of orthogonal tRNAs identified seemed to indicate that *E. coli* synthetases are relatively specific for their substrate and cannot accept tRNAs which differ at some key positions, hence I was expecting a very large

number of the tested synthetases to display a similar selectivity for their substrate and not to recognise the *E. coli* tRNAs. This result could potentially be explained by the arbitrary condition under which the *in vitro* test was performed, however, being unable to design an experiment capable of producing unambiguous results indicating the orthogonality *in vivo* of the active synthetases, I could only deduced that the *Ca*-ArgRS, *Ap*-HisRS and *Af*-TyrRS are orthogonal.

The application of an orthogonal pair to genetic code expansion most of the times requires to re-direct a tRNA to a codon different from its wild type codon, such as the amber stop codon,. Since the anticodon represent a crucial interaction point for most pairs, single point mutations in that region are enough to completely disrupt their interaction. In our case, we identified orthogonal pairs composed of an active synthetase and a tRNA with its native anticodon. We hence decided to verify how the identified pairs behaved upon conversion of the tRNAs to amber suppressors. Given the previous considerations, we were expecting some of the mutant tRNAs to not be recognised any longer from their cognate synthetases. However, since we had previously verified the activity of the wild type pair, we could pin down the loss of interaction to the mutation in the tRNA's anticodon and use molecular engineering to evolve the synthetase to accept its mutant tRNA as a substrate. Hypothetically, we could have immediately mutated the tRNAs to amber suppressor and use amber suppression to test orthogonality of these tRNAs and activity of their cognate synthetases. However, this approach would have resulted in a large number of inactive pairs for which no conclusion could be drawn. Hence, while producing pairs not immediately applicable to genetic code expansion, our decision to verify the wild type pairs provided us with a clear interpretation of our results.

Importantly, since mutations of the tRNAs and of the synthetases could alter the orthogonality of both, I decided to produce amber suppressors also for pairs for which the enzyme did not seem orthogonal based on the *in vitro* data. In the next chapter I will describe my efforts to produce active and orthogonal amber suppressors derived from the active pairs identified as described in this chapter.

Chapter III – Evolution of Orthogonal Aminoacyl-tRNA Synthetase/tRNA Pairs

Generation of Amber Suppressors

In the previous chapter I have described how tREX allowed me to experimentally characterise pairs composed of an orthogonal tRNA and an active synthetase (**Figure 2.8**). However, subsequent experiments aimed at verifying orthogonality for the synthetases resulted in the identification of fewer orthogonal enzymes than expected (**Figure 2.9**). Since the characterisation had been performed on wild type tRNAs with their native anticodon, the identified pairs could used as a starting point to reassign the corresponding sense codons to a new amino acid. In practice, however, in spite of the recent progress made for genetic codon reassignment and genome recoding, the use of sense codons for genetic code expansion is seldom a feasible option in most model organisms, since decoding the sense codons with non-canonical amino acids would poison the proteome and interfere with the

cellular functions. In practice, then, I acknowledged that, in the time frame of my doctoral research, applications of these pairs would still have to rely on conventional methods, such as amber suppression.

Engineering of a pair to be an effective amber suppressor, though, should not be considered straightforward. In fact, mutation of the tRNA's anticodon to CUA, the amber suppressor anticodon, generates a tRNA which can lose the features of its parent tRNA, such as orthogonality. In addition, since all of the identified active pair were composed of a synthetases which directly contacted the tRNA's anticodon, the conversion was expected to interfere with the recognition of the tRNA by the enzymes. Importantly, however, since all the pairs were know to be active, all the observed difference in behaviour would be directly attributed to the specific alteration of the anticodon and could be tackled by means of molecular evolution. In particular, in case of loss of orthogonality of the tRNAs, introduction of additional point mutations would be attempted to abolish the newly formed recognition by an endogenous *E. coli*. In case of loss of activity of the aaRS on the mutant tRNAs, mutation of those residues responsible for the recognition of the anticodon could be a solution to restore recognition in the mutant variant of its substrate. The generation of mutant synthetases capable of establishing specific interactions with the anti-amber anticodon, which is not present in any of the endogenous *E. coli* tRNAs, might have the additional benefit of increasing orthogonality for synthetases lacking a strong discrimination for their cognate tRNA. In light of this considerations, I decided that I would also attempt to generate an amber suppressing pair from pairs whose synthetases did not perform satisfactorily in the *in vitro* aminoacylation assay.

I selected nine distinct tRNAs, each belonging from a different isoacceptor class, as a starting point to generate amber suppressors:

(1) $tRNA^{Arg}_{GCG}$ from *Caldilinea aerophila* (*Ca*− $tRNA^{Arg}_{GCG}$);

(2) $tRNA^{Asp}_{GUC}$ from *Sorangium cellulosum* (*Sc*− $tRNA^{Asp}_{GUC}$);

(3) $tRNA^{Cys}_{GCA}$ from *Moorea producens* (*Mp*− $tRNA^{Cys}_{GCA}$);

(4) $tRNA^{Gln}_{UUG}$ from *Ilumatobacter nonamiensis* (*In*− $tRNA^{Gln}_{UUG}$);

(5) $tRNA^{Glu}_{CUC}$ from *Sporolactobacillus inulinus* (*Si*− $tRNA^{Glu}_{CUC}$);

(6) $tRNA^{Gly}_{CCC}$ from *Bacteroides vulgatus* (*Bv*− $tRNA^{Gly}_{CCC}$);

(7) $tRNA^{His}_{GUG}$ from *Afifella pfennigii* (*Ap*− $tRNA^{His}_{GUG}$);

(8) $tRNA^{Ile}_{UAU}$ from *Chitinophaga pinensis* (*Cp*− $tRNA^{Ile}_{UAU}$);

(9) $tRNA^{Tyr}_{GUA}$ from *Archaeoglobus fulgidus* (*Af*− $tRNA^{Tyr}_{GUA}$).

I started by mutating the anticodon of each of those tRNAs to CUA, without additional mutation to the anticodon arm nor any other part of the tRNA body. I cloned the suppressor tRNAs in plasmids either in absence or presence of their cognate synthetase. Unfortunately, generation of plasmids containing the mutant tRNA *(8)* $Cp-\mathrm{tRNA}^{\mathrm{Ile}}_{\mathrm{CUA}}$ failed repeatedly, hence this tRNA was not investigated further. The plasmids containing all the other pairs were challenged in an amber suppression assay by using a commonly used antibiotic resistance reporter, i.e.: the chloramphenicol acetyl-transferase gene interrupted at position 112 by a stop codon (*cat*^{112*}). *E. coli* DH10b cells containing the reporter were transformed with the plasmids containing the tRNAs alone or the tRNAs together with their cognate synthetases. The resulting transformants were tested for their level of resistance to the antibiotic chloramphenicol by verifying their ability to grow on plates containing either 0, 25, 50, 75, 100, 125, 150, 200 or 250 μg/mL of antibiotic. The results summarised below indicate the highest concentration of antibiotic tolerated by the cells (**Table 3.1**). It is important to notice that, while tREX represents a direct measurement of aminoacylation, reporter read-through is an indirect measurement which is additionally influenced by how efficiently an aminoacylated tRNA can take part in ribosomal translation. Consequently, the measurement of orthogonality by these two means are not directly comparable. The experiment summarised below and the discussion that follows use read-through as a proxy for aminoacylation and orthogonality.

		(-) aaRS	(+) aaRS			(-) aaRS	(+) aaRS
(1)	$Ca-\mathrm{tRNA}^{\mathrm{Arg}}_{\mathrm{CUA}}$	250	250	*(5)*	$Si-\mathrm{tRNA}^{\mathrm{Glu}}_{\mathrm{CUA}}$	0	0
(2)	$Sc-\mathrm{tRNA}^{\mathrm{Asp}}_{\mathrm{CUA}}$	0	0	*(6)*	$Bv-\mathrm{tRNA}^{\mathrm{Gly}}_{\mathrm{CUA}}$	25	25
(3)	$Mp-\mathrm{tRNA}^{\mathrm{Cys}}_{\mathrm{CUA}}$	75	250	*(7)*	$Ap-\mathrm{tRNA}^{\mathrm{His}}_{\mathrm{CUA}}$	250	250
(4)	$In-\mathrm{tRNA}^{\mathrm{Gln}}_{\mathrm{CUA}}$	0	0	*(9)*	$Af-\mathrm{tRNA}^{\mathrm{Tyr}}_{\mathrm{CUA}}$	250	250

Table 3.1*: Highest chloramphenicol concentration tolerated, among the ones tested, by E. coli cells due to the read-through of the cat*^{112*} *reporter gene by the amber suppressor tRNA indicated, in the presence or absence of its cognate aaRS.*

The results shown above highlighted how $Sc-\mathrm{tRNA}^{\mathrm{Asp}}_{\mathrm{CUA}}$, $In-\mathrm{tRNA}^{\mathrm{Gln}}_{\mathrm{CUA}}$ and $Si-\mathrm{tRNA}^{\mathrm{Glu}}_{\mathrm{CUA}}$ retained complete orthogonality as their parent tRNAs, however, they completely lost their ability to productively interact with their partner synthetase. This result was fully in line with the considerations which were made before. $Bv-\mathrm{tRNA}^{\mathrm{Gly}}_{\mathrm{CUA}}$ displayed only a very minor background aminoacylation by endogenous synthetases, but did nor appear to be recognised to any extent by its cognate synthetase either. $Mp-\mathrm{tRNA}^{\mathrm{Cys}}_{\mathrm{CUA}}$ displayed a modest background level of aminoacylation, suggesting that it had partially lost its orthogonality following alteration of its anticodon, however the reporter production increased significantly in the presence of its cognate synthetase. $Ca-\mathrm{tRNA}^{\mathrm{Arg}}_{\mathrm{CUA}}$, $Ap-\mathrm{tRNA}^{\mathrm{His}}_{\mathrm{CUA}}$ and $Af-\mathrm{tRNA}^{\mathrm{Tyr}}_{\mathrm{CUA}}$ lost their orthogonality and their aminoacylation level did not appear to be modulated by the presence or absence of their cognate synthetase.

The levels of chloramphenicol resistance of some of the samples described above clearly indicated that the newly generated amber suppressor tRNAs were not always orthogonal. Given the absence of amber suppressors in *E. coli* DH10b, a useful way to characterise which of the endogenous *E. coli* aaRSs was responsible for this mis-aminoacylation consisted in the determination of which amino acid was inserted in a target protein in response to an amber stop codon in its message RNA. In our lab and others this analysis was commonly performed by using sfGFP[121] as a reporter protein, due to

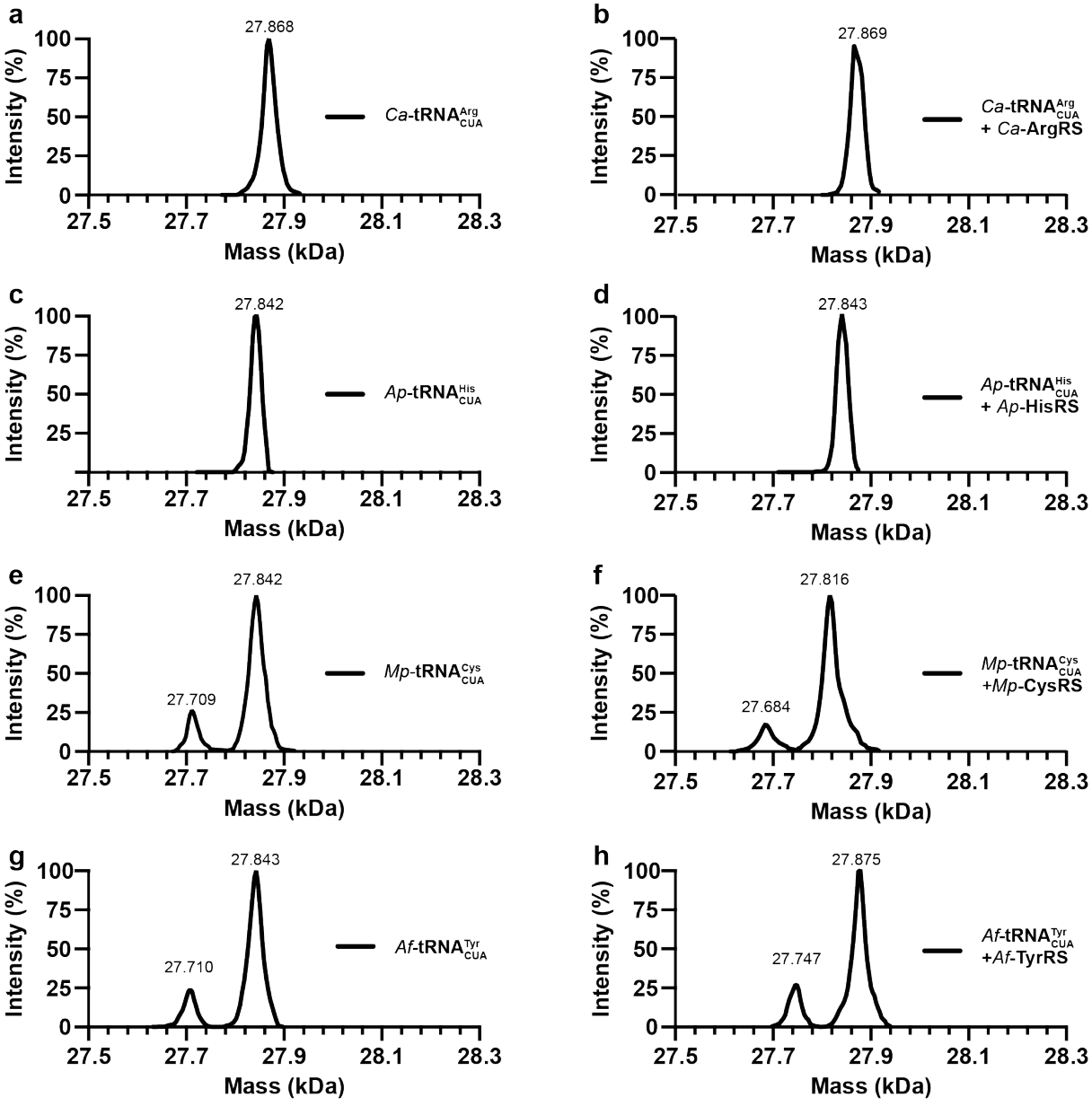


***Figure 3.1**: LC-MC analysis of the sfGFP[150]* reporter purified from cells expressing **a)** Ca−tRNA$^{Arg}_{CUA}$; **b)** Ca−tRNA$^{Arg}_{CUA}$ +Ca-ArgRS; **c)** Ap−tRNA$^{His}_{CUA}$; **d)** Ap−tRNA$^{His}_{CUA}$ +Ap-HisRS; **e)** Mp−tRNA$^{Cys}_{CUA}$; **f)** Mp−tRNA$^{Cys}_{CUA}$ +Mp-CysRS; **g)** Af−tRNA$^{Tyr}_{CUA}$; **h)** Af−tRNA$^{Tyr}_{CUA}$ +Af-TyrRS. The average masses detected are indicated above each peak. The minor peaks correspond to the removal of Met1 from the mature protein. Results are summarised in **Table 3.2**. No additional peaks where detected within the detection limit of the instrument (~1 % of the total protein injected), so that a minor amount of other species might be present.*

its high expression levels, exceptional stability, ease of purification and small size, which makes it suitable of mass spectrometry. The position at which the amber stop codon was inserted corresponds to the first amino acid of the 7th β-strand facing outward from the barrel, equivalent to position 150 in our particular variant (*gfp*^{150}*), as this position has been tolerant to all of the ncAAs tested in the lab. Given the sensitivity of mass determination by the instrument, mass spectrometry on the purified protein is suitable to discriminate the amino acids which differ in mass by more than 1 Da, i.e. all the amino acids except the pairs ile-leu, asn-asp and the group gln-glu-lys, as their mass is too similar.

LC-MC analysis of the sfGFP^{150}* reporter following amber suppression using the tRNAs under investigation (**Figure 3.1**) revealed aminoacylation by the following amino acids:

		(-) aaRS	(+) aaRS			(-) aaRS	(+) aaRS
(1)	*Ca*− $tRNA^{Arg}_{CUA}$	arg	arg	*(3)*	*Mp*− $tRNA^{Cys}_{CUA}$	gln/glu/lys	cys
(7)	*Ap*− $tRNA^{His}_{CUA}$	gln/glu/lys	gln/glu/lys	*(9)*	*Af*− $tRNA^{Tyr}_{CUA}$	gln/glu/lys	tyr

Table 3.2*: amino acid charged onto the amber suppressor tRNA indicated, either in the presence or the absence of its cognate synthetase, as indicated by the mass spectrometry measurement of the total mass of the sfGFP^{150*} reporter. Reference masses are indicated in* ***Chapter IV – Materials & Methods: GFP Total Mass****.*

Interpreting the information contained in **Table 3.1** and **Table 3.2** together, I drew the following conclusions:

(1) *Ca*− $tRNA^{Arg}_{CUA}$ was aminoacylated efficiently by the endogenous *E. coli* ArgRS, which led to a significant increase to chloramphenicol resistance. Experimental evidences were not enough to support nor disprove the activity of the *Ca*-ArgRS on the amber suppressing mutant of its tRNA. The presence of a background incorporation with the same amino acid as the natural substrate for the synthetase complicated the possibility to engineer the pair to improve orthogonality. In fact, in case the *Ca*-ArgRS could not recognise the *Ca*− $tRNA^{Arg}_{CUA}$, the evolution of the enzyme required to re-establish their interaction would be made difficult by the fact that the tRNA would be aminoacylated in any case with arginine. Additionally, reduction of the background recognition by the *Ec*-ArgRS would require mutation of the tRNA, whose effect on the interaction with the *Ca*-ArgRS could not be monitored for similar reasons. Overall, these considerations led me to not pursue the goal of obtaining an efficient amber suppressor pair from this tRNA/synthetase;

(2) *Sc*− $tRNA^{Asp}_{CUA}$ had completely lost its ability to be recognised by the *Sc*-AspRS, but retained full orthogonality. This confirmed that recognition of the anticodon represented a key feature of the interaction and suggested that engineering the anticodon-binding domain of the synthetase might restore it;

(3) *Mp*− $tRNA^{Cys}_{CUA}$ displayed a moderate background, but a robust incorporation of cysteine was

driven by the concomitant presence of *Mp*-CysRS in the cells. In this case, molecular evolution of either the synthetase or the tRNA might lead to an amber suppressor pair with improved performances;

(4) *In*− $tRNA^{Gln}_{CUA}$ behaved like *Sc*− $tRNA^{Asp}_{CUA}$ and had completely lost its ability to be recognised by the *In*-GlnRS, while still being fully orthogonal. This result was unexpected, considering that other glutaminyl pairs, notably the one from *E. coli*, were known to be permissive to conversion to amber suppressors[122], and suggested that the *In*-GlnRS had evolved a distinct modality of interaction with its cognate tRNA, with a stronger emphasis on the recognition of the anticodon compared to its orthologues from other species. Also in this case, engineering of the synthetase in its anticodon-binding domain was considered a valuable strategy to restore the interaction;

(5) *Si*− $tRNA^{Glu}_{CUA}$ followed a similar pattern as *Sc*− $tRNA^{Asp}_{CUA}$ and engineering of the enzyme was considered a valuable option to restore interaction between the tRNA and the synthetase;

(6) *Bv*− $tRNA^{Gly}_{CUA}$ displayed only a minor level of background aminoacylation, however its recognition by the cognate synthetase was completely abolished. Given the existence of extensive interaction between the enzyme and the tRNA, we reasoned that evolution of the anticodon-binding domain of the synthetase could potentially allow us to identify a variant which can engage the mutant anticodon in a new interaction network;

(7) *Ap*− $tRNA^{His}_{CUA}$ lost its orthogonality compared to its wild type counterpart, however, unlike the *Ca*− $tRNA^{Arg}_{CUA}$, it became a substrate for one of the *E. coli* synthetase (either the *Ec*-GlnRS, *Ec*-GluRS, or the *Ec*-LysRS) from a different isoacceptor class. Furthermore, the experiments highlighted unambiguously that the synthetase did not display any activity on the mutant tRNA. We reasoned though that engineering of the synthetase might restore the interaction and that this would lead to histidine being charged onto the tRNA, which could be monitored by mass spectrometry;

(8) *Af*− $tRNA^{Tyr}_{CUA}$ lost its orthogonality following the single point mutation G34C in its anticodon. As a consequence, the mutant tRNA became a substrate for either *Ec*-GlnRS, *Ec*-GluRS, or the *Ec*-LysRS. The levels of chloramphenicol resistance observed in the presence of the *Af*-TyrRS did not differ from the ones observed in its absence, however mass spectrometry indicated a qualitative difference in the incorporation which changed from gln/glu/lys to tyr, confirming some level of activity of the synthetase towards its mutant tRNA. The absence of background aminoacylation (up to detection limit of the instrument) in the GFP purified from cells containing both the *Af*− $tRNA^{Tyr}_{CUA}$ and the *Af*-TyrRS indicated that competition can effectively alter the orthogonality of a tRNA, i.e.: the *Af*− $tRNA^{Tyr}_{CUA}$ was orthogonal to the

endogenous *E. coli* synthetases in the presence of *Af*-TyrRS but not in its absence. This same observation also implied that the *E. coli* synthetase which mis-charged the $Af-\mathrm{tRNA}_{\mathrm{CUA}}^{\mathrm{Tyr}}$ was orthogonal to that tRNA in the presence of *Af*-TyrRS but not in its absence. This experimental evidence further cast a doubt on the validity of *in vitro* aminoacylation assays in the assessment of synthetases orthogonality. In order to improve the effectiveness of the pair as an amber suppressor, we decided to both try to engineer the synthetase to optimise its interaction with the mutant tRNA and to mutate the tRNA to improve its orthogonality in the absence of its cognate synthetase.

In the following sections I will describe the experimental procedure which was followed to try and engineer the pairs as described herein.

tRNA^{Asp} from *Sorangium cellulosum*

The aspartyl-tRNA synthetase from *S. cellulosum* (*Sc*) is a dimeric enzyme belonging to the class II of aaRS. Although a structure of this protein is not available, the crystal structure of its orthologue from *Thermus thermophilus* (*Tt*) (PDB 1efw[123], **Figure 3.2a**) allowed to determine that the enzyme is composed of a small N-terminal domain (**Figure 3.2a**, blue), encompassing approximately the initial 20% of its primary sequence, which contacts the anticodon of the tRNA by well defined hydrogen bonds to positions 34, 35 and 36; and a C-terminal domain, responsible for the dimerisation and for the catalysis itself.

While the experimental evidence (**Figure 2.9**, Asp_09) indicated that the *Sc*-AspRS was active on the *Sc*− $tRNA^{Asp}_{GUC}$, amber suppression experiments on *Sc*− $tRNA^{Asp}_{CUA}$ in the presence of *Sc*-AspRS did not reveal any chloramphenicol resistance even at the lowest concentration tested. I concluded that the two mutations G34C and C36A had completely impaired the recognition between the two partners. This could be due to either the loss of a fundamental fraction of the binding energy of the pair, which could be provided by the hydrogen bonds formed between the anticodon and the enzyme, or by a steric clash caused by the additional hinderance caused by the transversion C36A. The crystal (**Figure 3.2b**) reveals that only few residues are responsible for the establishment of the hydrogen bond network between the synthetase and the tRNA's anticodon: in particular, E91 in the *Tt*-AspRS contacts the guanidine moiety of the modified G34 nucleotide of the tRNA, R78 interacted with the carbonyl groups of both U35 and C36, while the latter was also engaged by the amide group of N82. Notably, the majority of the interactions between the side chains from the enzyme and the nucleobases from the tRNA are constituted by charge-to-dipole interactions, which are stronger then dipole-to-dipole interactions, thus potentially providing a significant contribution to the binding energy responsible to keep the partners together during aminoacylation.

Conversion of the *Sc*− $tRNA^{Asp}_{GUC}$ to an amber suppression did not require mutations of the middle position of the anticodon, however, since the only residue which interacted with U35 (i.e.: Arg78) was also responsible for the recognition of C36, which in turn was mutated to a bulkier A36, I decided to generate a library of variants of the *Sc*-AspRS which would randomise the residues of the enzyme corresponding to Arg78, Asn82 and Glu91 from the *Tt*-AspRS (**Figure 3.2b**). Alignment of the primary sequence of the two proteins revealed that these residues were Asn107, Arg102 and Glu116 from *Sc*-AspRS, which were then randomised to all possible natural amino acids by enzymatic inverse PCR

(eiPCR[124]) to generate a library of mutant synthetases. The library was generated using a plasmid containing both the synthetase and its cognate tRNA as a starting material. In order to verify whether any of the generated variants could re-establish any productive interaction leading to aminoacylation of the tRNA, I transformed the library into a reporter strain of *E. coli* DH10b constitutively expressing the *cat*^{112*} reporter, which conferred resistance to chloramphenicol if the cell could perform amber

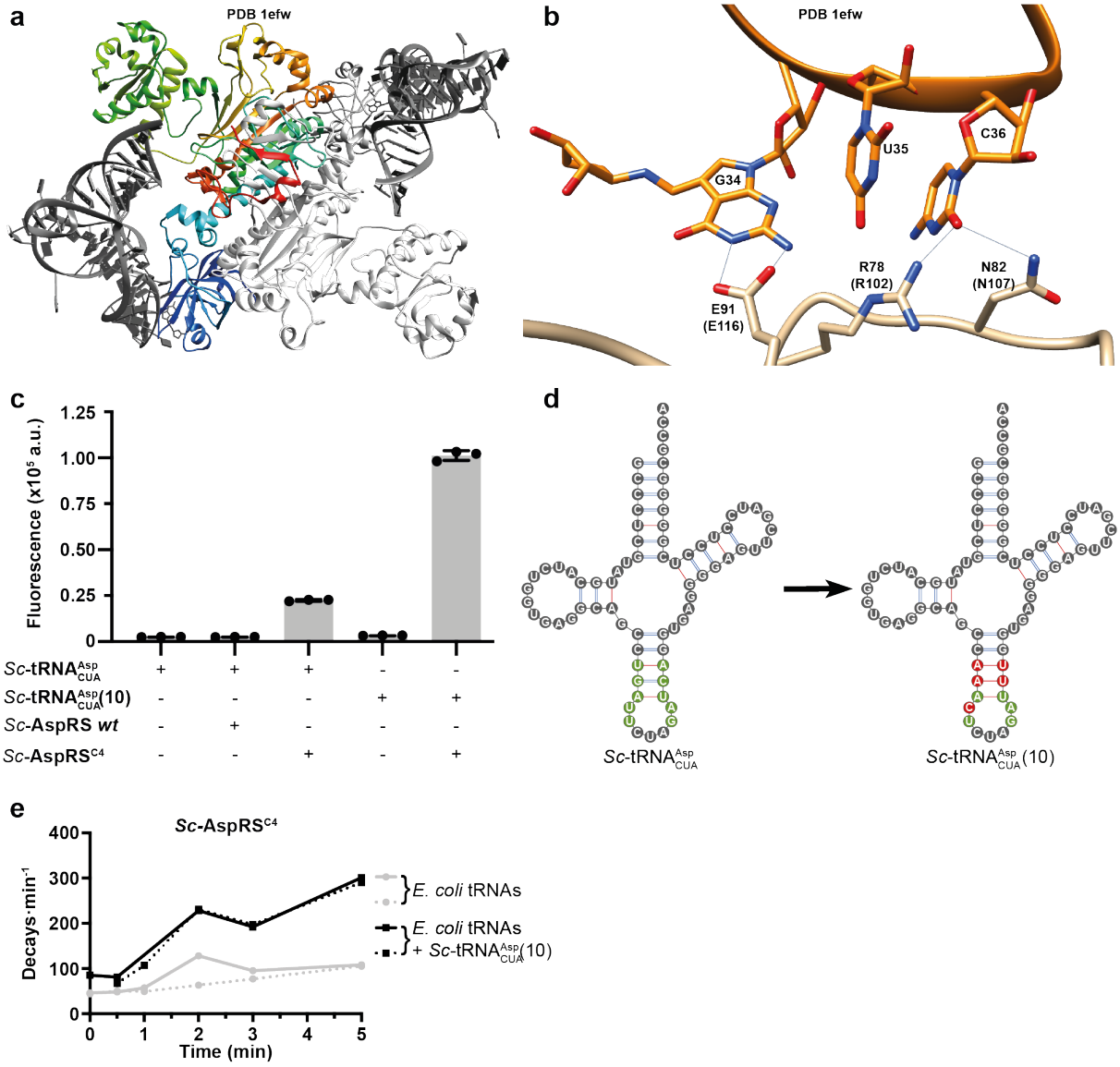


***Figure 3.2**: Evolution of the S. cellulosum tRNA^{Asp}/AspRS pair. **a)** Crystal structure of the closest orthologue of the Sc-AspRS from T. thermophilus (PDB 1efw[123]). The enzyme assembles into dimers, one subunit of which is coloured as a rainbow (N-terminal blue to C-terminal red) while the second subunit being shown in white, for clarity. tRNAs are coloured in dark grey. **b)** Interaction between the tRNA anticodon (orange) and the residues at the N-terminal domain of the synthetase. **c)** GFP expression level of various combination of tRNA^{Asp}/AspRS during the evolution of the pair. **d)** Schematic structure of the amber suppressor mutant of Sc-tRNA^{Asp}. The residues in green were randomised in a library from which the mutant Sc−tRNA$^{Asp}_{CUA}$ (10) was selected. The mutated residues are shown in red. **e)** In vitro aminoacylation experiment using the evolved Sc-AspRS^{C4} showed a distinct aminoacylation rate in the absence or presence of a specific substrate for the synthetase, supporting its orthogonality.*

suppression effectively, and the *gfp*^{150*} reporter upon induction of the L-arabinose inducible P_{BAD} promoter, which made the cells visibly fluorescent when illuminated using visible blue light (488 nm).

The two reporters, expressed from the same reporter plasmid called p15A-*cat*^{112*}-*gfp*^{150*}, served two distinct purposes: the antibiotic resistance marker allowed the survival on plates of only the cells harbouring an active synthetase with restored activity for its cognate tRNA, reducing drastically the number of cells capable of forming colonies on solid medium.. The GFP provided a fast dynamic measurement of the efficiency of amber suppression by the pair. Furthermore, I knew from experimental experience that positive selections on chloramphenicol are permissive to some extent to the survival of clones with low or absent amber suppression activity, while the incidence of false positive clones displaying both antibiotic resistance and green fluorescence in the absence of effective amber suppression is mostly negligible.

A round of positive selection using the dual reporter cell line grown on plates containing low concentrations of chloramphenicol (i.e.: 50 ng/mL, chosen on the basis of the results shown in **Table 3.1**) allowed me to identify a clone possessing moderate but reliably reproducible amber suppression properties (**Figure 3.2c**). This cloned, which was called *Sc*-AspRS^{C4} presented only two mutations: Arg102Asn and Glu116Arg, while retaining the wild type Asn107. I speculated that the reason why these mutations were sufficient to restore the interaction between the $Sc-tRNA_{CUA}^{Asp}$ and the *Sc*-AspRS was the generation of a new Arg116:C34 charge-to-dipole interaction which had replaced the native Glu116:G34 interaction, since the longer side chain of arginine compared to glutamic acid could compensate for the smaller size of cytosine compared to guanine, while also presenting the appropriate complementary pattern of hydrogen bonds. Additionally, the minor steric bulk of Asn102 compared to Arg102 could allow the accommodation of the larger A36 which had taken the place of C36, while also allowing for the formation of potential hydrogen bonds.

The identification of the mutant described was very promising, however the overall efficiency of the engineered pair as an amber suppressor remained modest, as measured by the GFP fluorescence induced by the read-through of the *gfp*^{150*} reporter (**Figure 3.2c**). In order to verify whether this activity could be improved, I tested if the $Sc-tRNA_{CUA}^{Asp}$ could become a better substrate for the newly evolved *Sc*-AspRS^{C4}. In particular, given that the interface between the tRNA anticodon and the anticodon-binding domain of the synthetase had been entirely re-designed, I questioned whether mutation of the residues of the anticodon arm surrounding the anticodon could lead to improved performances. This approach, which had been used successfully before[41] has its foundation in two aspects: first, the anticodon arm had been observed to alter the efficiency of decoding of a specific codon by a tRNA[103]; secondly, while overall isosteric, different base pairs sequences are known to induce slightly different conformations to nucleic acids[125], such that mutations might potentially optimise at a finer level the surface complementarity between the tRNA and the synthetase. I had to

keep in mind, though, that any alteration to the tRNA could lead to alterations to its orthogonality, so that lack of aminoacylation in the absence of the cognate synthetase always had to be verified. To account for this, I altered the experimental design by generating a library of the 5 nt on either sides of the anticodon of *Sc*-$tRNA^{Asp}_{CUA}$ by eiPCR on a plasmid only containing the tRNA (**Figure 3.2d**). The gene coding for the *Sc*-AspRS^{C4} was cloned in the reporter plasmid p15A-*cat*^{112}*-*gfp*^{150}* to generate a new reporter known as p15A-*cat*^{112}*-*gfp*^{150}*-*ScAspRS*^{C4}. The library was transformed in cells harbouring the new reporter plasmid and the transformants were grown on selective conditions using a concentration of chloramphenicol (i.e.: 100 ng/mL) higher than the one used for the previous round of selection, in order to ensure that the surviving colonies would harbour a tRNA which would outperform its parental counterpart for amber suppression. The clones displaying the highest levels of GFP fluorescence were grown individually and the tRNA-containing plasmid was purified, then transformed in reporter cells containing either the p15A-*cat*^{112}*-*gfp*^{150}* or the p15A-*cat*^{112}*-*gfp*^{150}*-*ScAspRS*^{C4} plasmid. This experiment allowed me to verify the levels of GFP expression produced in the absence or presence of the *Sc*-AspRS^{C4}, thus examining both the tRNA orthogonality and the overall activity of the pair. This screening, performed together with Dr. Shan Tang, allowed me to identify a new tRNA, called *Sc*-$tRNA^{Asp}_{CUA}(10)$, which fully preserved orthogonality while displaying enhanced performance as an amber suppressor (**Figure 3.2c**). However, the experiments described did not allow me to verify whether the improvement derived from higher aminoacylation or better ribosomal decoding.

The evolution experiments described highlighted how the *Sc*-$tRNA^{Asp}_{CUA}$/*Sc*-AspRS pair was converted from a fully inactive pair to a good amber suppressing one by mean of molecular evolution. Furthermore the synthetase was evolved to actively recognise the CUA anticodon, which is not present in any of the endogenous *E. coli* tRNAs. For this reason we hypothesised that following evolution orthogonality of the pair might have been improved, as its discrimination for the anticodon had changed drastically. Consequently, I purified the *Sc*-AspRS^{C4} synthetase and tested it in an *in vitro* setting as I had done previously for its wild type counterpart. The new experiment (**Figure 3.2e**) showed a remarkable difference in the aminoacylation dynamics between the sample containing *Sc*-$tRNA^{Asp}_{CUA}(10)$ compared to the sample containing only the *E. coli* tRNAs, differently from what observed before the evolution (**Figure 2.9**). This observation led us to conclude that evolution had effectively generated a new orthogonal amber suppressing pair.

Engineering the Amino Acid Specificity for *Sc*-AspRS^{C4}

Since I could evolve the *S. cellulosum* tRNA^{Asp}/AspRS pair into an effective and orthogonal amber suppressing pair, I wondered whether I could alter its amino acid specificity in order to incorporate a ncAA for genetic code expansion. In order to understand how the synthetase recognised its amino acid substrate, I relied on the structure of the closest orthologue crystallised bound to aspartyl-AMP (**Figure 3.3a**, AspRS from *E. coli,* PDB 1c0a[126]). Analysis of the structure revealed that the side chain of the aspartic acid forms a salt bridge with a very conserved arginine residue (Arg489 in *E. coli* AspRS), which is in turn held in place by interactions with other side chains (Glu235 and Ser487) protruding from adjacent strands of the β-sheet which constitutes the catalytic pocket. In addition to the fundamental contribution provided from Arg489, Lys 198 established an additional salt bridge with the side chain of the substrate, while a pair of consecutive histidines at positions 495 and 496 form hydrogen bonds with the β-carboxyl group. In spite of its importance for substrate binding, Gln195 only interacts with the α-amine group and is for this reason not directly involved with the recognition of the side chain. Similarly to what could be observed in the crystal structure shown, previous studies on the eukaryotic enzyme from *S. cerevisiae*[127] had highlighted the conservation of the residues Arg489, Lys198 and Glu235 (Arg485, Lys306 and Glu344 in *S. cerevisiae,* respectively), while also showing how mutation of any of these residues led to complete inactivation of the enzyme.

From these pieces of information and from the alignment of the *S. cellulosum* AspRS to the *E. coli* orthologue I reasoned that Arg536 (corresponding to Arg489 in *E. coli*) had to play a pivotal role in the recognition of aspartic acid as a substrate for the enzyme, hence I decided to generate a library of mutant enzymes in which I randomised this position to all the possible canonical amino acids except for arginine.

The reasons behind this choice were firstly the likelihood that mutants retaining arginine at this position would most likely still recognise aspartic acid as their substrate, but also the observation that the long side chain of arginine occupies a very large volume of the catalytic pocket of the enzyme, such that exclusion of that residue from this position would generate a pocket of significantly larger size, potentially enabling incorporation of ncAAs with a bulkier side chain. In order to obtain this result with ease, I made use of a computational tool I developed in 2016, which expanded the generality of the approach described by Tang *et al.* in 2012 for the generation of so-called "small intelligent libraries"[128] by identifying which combination(s) of degenerate primers for site-saturation mutagenesis can be mixed in precise ratios to generate a mixture of sequence uniformly coding for a specific set of amino acids of choice. A complete description of this algorithm is available in my

Master's thesis (Daniele Cervettini, **Engineering the Methanogenic-Type Seryl-tRNA Synthetase**

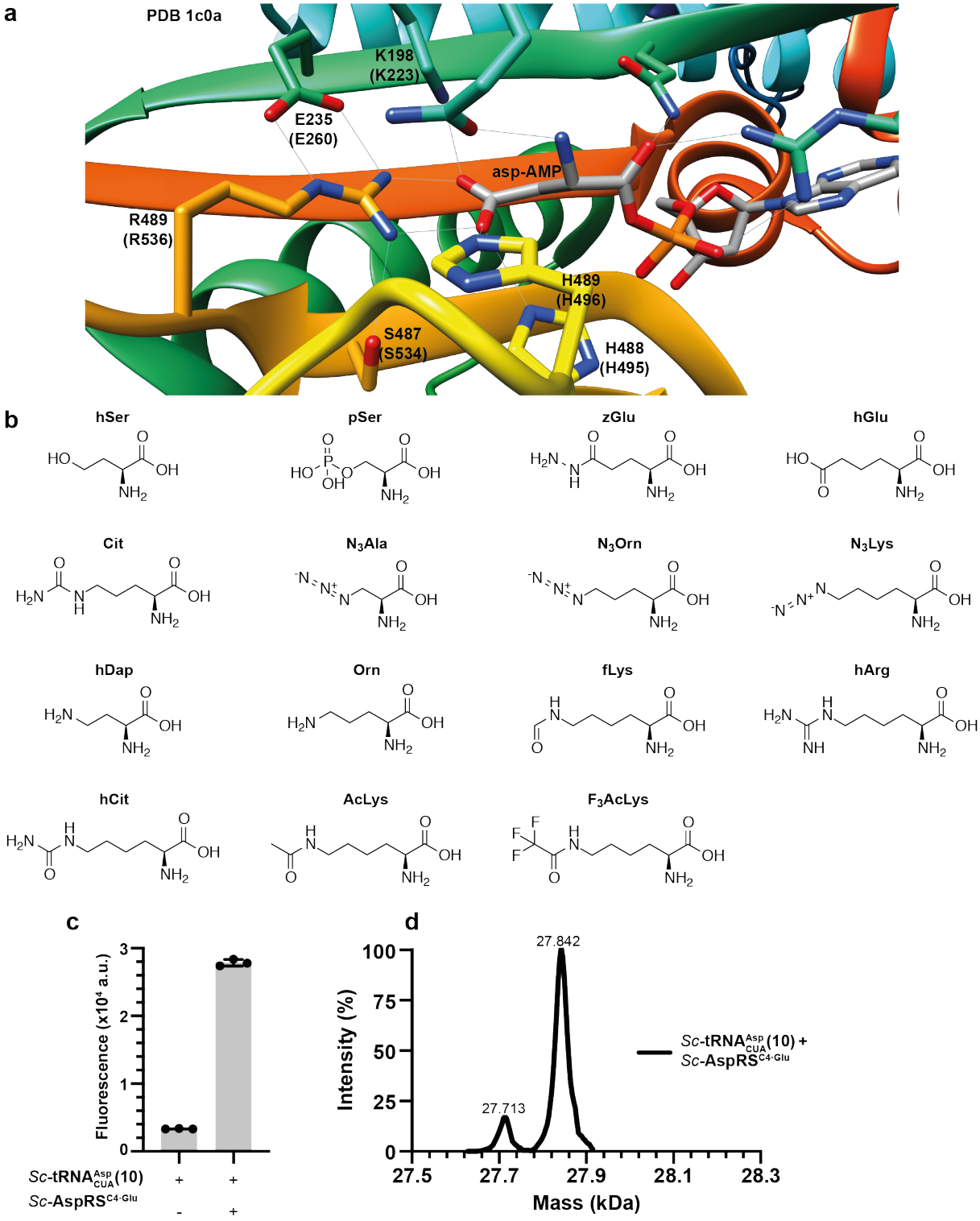


***Figure 3.3**: **a**) Molecular structure of the active site of the AspRS synthetase by E. coli (PDB 1c0a[126]) highlights the mechanisms of recognition of the substrate. **b**) Chemical structures of the ncAAs used for the selections of mutant Sc-AspRSs. **c**) Activity of the Sc-AspRS^{C4Glu} as measured by production of the sfGFP by read-through of the sfGFP^{150}* reporter. **d**) Mass spectrometry analysis on the purified sfGFP produced by the read-through of the sfGFP^{150}* reporter confirms incorporation of glutamic acid by the synthetase. The minor peak corresponds to the the cleavage of Met1 from the protein.*

from *Methanosarcina barkeri*: a New Methodological Approach, 2006, Università di Pisa), and the code is available in **Chapter V – Appendix: PrimDesign**.

In particular, while in their paper Tang *et al.* only described the combination of primers with sequence NDT:VMA:ATG:TGG to be mixed in ratios of 12:6:1:1 to obtain uniform coverage of all 20 amino acids, my software identified the combination VHG:WKC:NAC:KGG to be mixed in ratios of 9:4:4:2 to have uniform coverage of all the natural amino acids except for arginine.

In addition to position 536 above described, my library also randomised Lys223, Glu260 and Ser534 (corresponding to Lys198, Glu235 and Ser487 in the *E. coli* AspRS), respectively, to all possible amino acids (**Figure 3.3a**). Having hypothesised that the removal of the wild type arginine from position 536 would expand the size of the amino acid binding pocket, I decided to perform positive selection on the library against each of the ncAAs belonging to the set listed below, which included some polar amino acids of similar size compared to aspartic acid, but also some others with a larger side chain (**Figure 3.3b**):

1. homoserine (**hSer**, (*S*)-2-amino-4-hydroxybutanoic acid);
2. phosphoserine (**pSer**, (*S*)-2-amino-3-(phosphonooxy)propionic acid);
3. glutamic acid γ-hydrazide (**zGlu**, (2*S*)-2-amino-5-hydrazinyl-5-oxopentanoic acid);
4. homoglutamic acid (**hGlu**, (*S*)-2-aminohexanedioic acid);
5. citrulline (**Cit**, (*S*)-2-amino-5-(carbamoylamino)pentano acid);
6. azidoalanine (**N_3Ala**, (*S*)-2-amino-3-azidopropanoic acid);
7. azidoornithine (**N_3Orn**, (S)-2-amino-5-azidopentanoic acid);
8. azidolysine (**N_3Lys**, (S)-2-amino-5-azidohexanoic acid);
9. homoDAP (**hDap**, (*S*)-2,4-diamino-butanoic acid);
10. ornithine (**Orn**, (S)-2,5-diamino-pentanoic acid);
11. formyl-lysine (**fLys**, (S)-2-amino-6-formamidohexanoic acid);
12. homoarginine (**hArg**, (S)-2-amino-6-(diaminomethylideneamino)hexanoic acid);
13. homocitrulline (**hCit**, (S)-2-amino-6-(carbamoylamino)hexanoic acid);
14. acetyl-lysine (**AcLys**, (S)-6-acetamido-2-aminohexanoic acid);
15. trifluoroacetyl-lysine (**F_3AcLys**, (*S*)-2-amino-6-[(2,2,2-trifluoroacetyl)amino]hexanoic acid).

I set up a round of positive selection using the dual reporter system described in the previous paragraph for each for the 15 amino acids listed above at a concentration of 2 mM on solid medium.

Additionally, one positive selection plate was set up without the addition of any ncAA. After incubation at 37°C, a similar number of GFP-positive colonies were observed on each plate, including the negative control. Considering the special precautions taken in making the library, I hypothesises that the colonies observed could include some new mutants incorporating a different amino acid than aspartic acid. For this reason, I extracted the library plasmid contained in several clones grown on different plates in order to sequence the synthetase variant and to perform amber suppression on the sfGFP^{150*} reporter for mass spectrometry. The analysis reveal that clones retaining the wild type Lys223 and Glu260 but with the mutations Ser534Arg and Arg536Ala/Ser, or clones with the wild type Lys223 but with the mutations Arg536Tyr and with two small residues (Gly/Cys/Ser) at positions 260 and 534,. could retain a fraction of the activity of the wild type clone and the same specificity for aspartic acid, as confirmed by mass spectrometry. Interestingly, though, clones were identified which lacked the conserved Lys223 and had instead acquired the mutations (Lys223Ala, Glu260Lys and Arg536Ala/Asn/Gly, while retaining the wild type Ser534. These samples displayed modest but reproducible amber suppression activity, the most active variant being the mutant Lys223Ala, Glu260Lys and Arg536Ala, called *Sc*-AspRS$^{C4\ Glu}$ (**Figure 3.3c**), while mass spectrometry on these samples highlighted that the specificity of these mutants had changed to glutamic acid. No mutants were selected from this library that were able to incorporate any of the ncAAs tested.

While not representing the most desirable outcome, the identification of the *Sc*-AspRS$^{C4\ Glu}$ mutant proved that the active site of the enzyme can indeed be mutated to accommodate amino acids different from the wild type substrate of the synthetase. Even if only differing from aspartic acid by the presence of one extra carbon, glutamic acid seemed to be recognized by a distinct mechanism, as indicated by the loss of both the salt bridge-forming residues Lys223 and Arg536, while the negatively charged residue Glu260 had been replaced by the positively charged Lys260, likely responsible for the establishment of a new polar interaction pattern.

tRNA^{Cys} by *Moorea producens*

Cysteinyl-tRNA symthetases are a family of small monomeric enzymes belonging to the class I of aminoacyl-tRNA synthetases, lacking any editing domain due to their high selectivity for cysteine. The structure of the CysRS from *M. producens* has not been resolved before, however, the crystal structure of its orthologue from *E. coli* (**Figure 3.4a**, PDB 1u0b[129]) shows how the enzyme is composed of a catalytic N-terminal domain, encompassing about 85% of the whole primary sequence of the enzyme, while the C-terminal domain is composed of the last 15% of the protein and is responsible for the interactions with the tRNA's anticodon (**Figure 3.4a**, orange and red). In addition to specific interaction between the enzyme and the tRNA's anticodon, the structure shows a wide surface area of contact points between the two partners.

Conversion of the *Mp*-tRNA^{Cys} to an amber suppressor had not resulted in a significant loss of orthogonality, and furthermore the resulting amber suppressor could be still recognised by the wild type enzyme to induce a low but reproducible level of amber suppression of the sfGFP^{150*} (**Figure 3.4b**). The wild type anticodon for the *Mp*-tRNA^{Cys} is GCA, so that its conversion to an amber suppressor involved the two mutations G34C and C35U. Differently from the case of *Sc*-tRNA^{Asp}, neither of these mutations were transversion from pyrimidine to sterically bulkier purines, which would be more likely to disrupt interactions due to steric clashes. Furthermore, the persistence of interaction might have been facilitated by the broad contact area mentioned before. In spite of this, I wanted to verify whether the activity of the pair could be improved by re-optimising the interaction at the anticodon level.

A more in-depth analysis of the crystal structure of *E. coli* CysRS (**Figure 3.4c**) highlighted how Asp436 interacts specifically with the guanidine moiety of G34 in the anticodon, while its carbonyl group establishes a polar contact with the side chain of Arg427. Similarly, C35 is involved in a complex network of polar contacts with Arg439 and Asp451, the latter being held in place by a salt bridge formed with Arg423. Overall, position 34 and 35 of the anticodon are recognised by both positively and negatively charged residues, which form charge-to-dipole interactions. I generated a library of the residues Arg455, Arg459, Asp468, Arg471 and Asp483, which correspond to Arg423, Arg427, Asp436, Arg439 and Asp451 of the *E. coli* CysRS. This library underwent a round of positive selection using the dual reporter system described before and a concentration of chloramphenicol equal to 350 µg/mL due to the higher starting activity of the pair. Several clones among the ones

forming colonies under selective conditions displayed a high amber-suppression level, as estimate by observing the plates under a blue light (~488 nm). The most active 27 clones were isolated and the enzyme variant they contained was sequences. In this case, the active colonies did not display

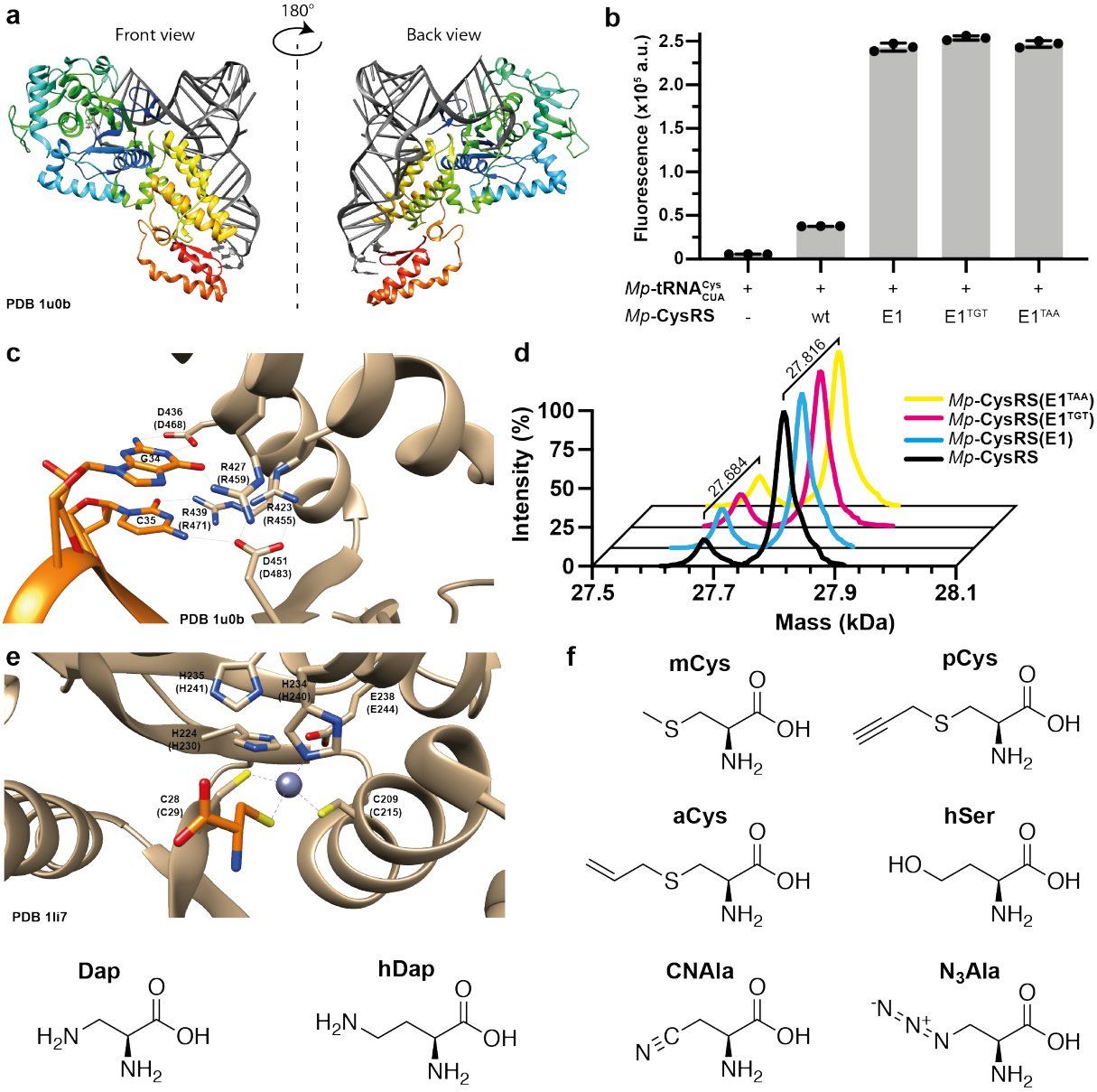


***Figure 3.4**: **a**) Crystal structure of the CysRS from E. coli (PDB 1u0b[129]) reveals the division of the enzyme into an N-terminal catalytic domain (blue to yellow) and a C-terminal anticodon-binding domain (orange and red). **b**) Activity of the amber suppressor mutant of the M. producens $tRNA^{Cys}$/CysRS pair, as measured by the read-through of the reporter sfGFP[150]*. The data indicated that while the amber suppressor tRNA retained orthogonality and while being capable of being recognised to a lesser extent by its wild type cognate CysRS, truncation of the C-terminal anticodon-binding domain significantly improved its performance. **c**) Molecular mechanism of the $tRNA^{Cys}$ anticodon recognition by the E. coli CysRS as revealed by the crystal structure of their complex (PDB 1u0b). **d**) Mass spectromery analysis of the read-through product of the sfGFP[150]* reporter reveals incorporation of cysteine on the $tRNA^{Cys}_{CUA}$ by all the variants of the synthetase tested. **e**) Molecular mechanism of the recognition of the amino acid substrate by the E coli CysRS (PDB 1li7[130]) reveals the role of the Zn^{2+} ion in the active site. **f**) Chemical structure of the ncAA used for the selection of mutant Mp-CysRS variants.*

sequence convergence, but 11 of them showed a TAG stop codon at either position 455, 459 or 471, including the clone E1 which showed a remarkable activity (**Figure 3.4b**) and harboured the mutations Arg455(TAG), Arg459Gln, Asp468Gly, Arg471Gln and Asp483Thr.

The presence of an amber stop codon at the first randomised position of the library raised a question. In fact, as the cells containing the *Mp*-CysRS^{E1} mutant displayed high amber suppression activity, I could not conclude whether the activity of the synthetase was resulting from the production of a truncated enzyme, given by termination at position 455, lacking almost the entire C-terminal anticodon-binding domain, or whether the active variant would result form suppression of the amber stop codon at position 455 using cysteine. In order to verify which one of the two options was correct, I mutated the sequence of the codon 455 from TAG to either the TGT cysteine codon, or to the TAA stop codon which cannot be suppressed, such that the two proteins would either contain or lack the C-terminal domain. The two variants, called *Mp*-CysRS$^{E1(TGT)}$ and *Mp*-CysRS$^{E1(TAA)}$ were tested, together with the *Mp*-CysRS^{E1} mutant containing the amber stop codon. The read-through activity of all the variants tested resulted completely equivalent (**Figure 3.4b**), and mass spectrometry analysis on the sfGFP produced in the assay confirmed the incorporation of cysteine at position 150 from all the variants tested, namely, the wild type *Mp*-CysRS and the mutants *Mp*-CysRS^{E1}, *Mp*-CysRS$^{E1(TGT)}$ and *Mp*-CysRS$^{E1(TAA)}$ (**Figure 3.4d**). These results suggested that the C-terminal domain of the enzyme was actually of hinderance for the optimal activity of the synthetase on the suppressor tRNA and that it was completely disposable. Given that the activity of the truncated mutant was indistinguishable from the mutant containing the amber stop codon and also from the mutant containing cysteine, I hypothesised that the mutant *Mp*-CysRS$^{E1(TGT)}$ could not form a functionally folded C-terminal domain, hence the equivalence to the *Mp*-CysRS$^{E1(TAA)}$ mutant.

Having generated a very active amber suppressor pair, I wanted to verify whether it could be used to incorporate ncAAs, hence I used the structural information of the *E. coli* CysRS crystallised bound to cysteine (**Figure 3.4e**, PDB 1li7[130]) to identify the residues responsible for the recognition of the amino acid. Interestingly, the structure highlights how the active site of the enzyme contains a Zn^{2+} ion which is coordinated by Cys28, Cys209, His234 and Glu238 in the absence of the amino acid. Upon binding, the thiol group of the free cysteine replaces Glu238 in the coordination of the Zn^{2+} ion. This recognition mechanism ensures the specificity of the enzyme without requiring the presence of an additional editing domain. I reasoned that incorporation of any ncAAs would require the removal of the Zn^{2+} ion from the active site, which would generate a new catalytic pocket of bigger size and perhaps allow incorporation of amino acids with bulkier side chains. I identified the residues Cys29, Cys215, His240 and Glu244 in *Mp*-CysRS as the ones responsible for the coordination of the Zn^{2+} ion. To minimise the background, I decided to exclude cysteine from the randomisation at positions 29 and 215, to be sure that the wild type coordination shell of the metal would be destroyed. To do so, I used my script to identify the combination of degenerate primers with sequence SVT, WHC, VWG

and TGG to be mixed in a ratio of 6:6:6:1 which would produce uniform coverage of the other 19 amino acids excluding cysteine. Additionally, I decided to randomise residues His230 and His241 as they are in close proximity to the active site. I decided to screen the library against the following set of amino acids (**Figure 3.4f**):

1. *S*-methyl-cysteine (**mCys**, (*R*)-2-amino-3-(methylmercapto)propionic acid);
2. *S*-propargyl-cysteine (**pCys**, (*R*)-2-amino-3-prop-2-ynylsulfanylpropanoic acid);
3. *S*-allyl-cysteine (**aCys**, (*R*)-2-amino-3-prop-2-enylsulfanylpropanoic acid);
4. homoserine (**hSer**, (*S*)-2-amino-4-hydroxybutanoic acid);
5. DAP (**Dap**, (*S*)-2,3-diaminopropanoic acid);
6. homoDAP (**hDap**, (*S*)-2,4-diamino-butanoic acid);
7. cyanoalanine (**CNAla**. (*S*)-2-Amino-3-cyanopropanoic acid);
8. azidoalanine (**N_3Ala**, (*S*)-2-amino-3-azidopropanoic acid).

Unfortunately, none of these selections revealed any mutant synthetase capable of incorporating alternative substrates. This undesired result might be due to different factors. First of all, it might be possible that the loss of the Zn^{2+} ion, required to fit any alternative substrate within the active site of the enzyme, might destabilise the folding of the protein, resulting in inactive variants. Alternatively, a different selection of ncAAs and/or a different library design might result in successful evolution of the pair.

tRNA^{Gln} from *Ilumatobacter nonamiensis*

Glutamine-tRNA synthetases are monomeric class I synthetases lacking an editing domain. The crystal structure of the *E. coli* GlnRS bound to its tRNA (**Figure 3.5a**, PDB 1qrs[131]) highlights how, differently from other synthetases, the enzyme is not composed of two distinct domains, one of which is responsible for the catalysis while the other being responsible for interaction with the tRNA's anticodon. Instead, the enzyme folds in a single globular structure which makes contacts with the tRNA on several points along its sequence. Importantly, the *E. coli* GlnRS is known to tolerate the mutant amber suppressor $Ec-\text{tRNA}^{\text{Gln}}_{\text{CUA}}$ which arises from the mutation of the *E. coli* $Ec-\text{tRNA}^{\text{Gln}}_{\text{CUG}}$ [122].

In contrast to the case just mentioned, conversion of the $In-\text{tRNA}^{\text{Gln}}_{\text{UUG}}$ to an amber suppressor did not result in any appreciable aminoacylation of the resulting tRNA, as assessed by the complete lack of read-through of the sfGFP^{150*} reporter (**Figure 3.5b**). This information highlighted how the tRNA recognition mechanism employed by the two enzymes should differ significantly, in spite of the homology between the to proteins. Due to the existence of a tRNA^{Gln} with anticodon CUG in *I. nonamiensis*, I concluded that the loss of recognition of the amber suppressor tRNA would have to be dependent on the G36A mutation. From the crystal structure of the *E. coli* GlnRS I identified the residues composing the loop Lys398, Gln399, Tyr400, Lys401 and Arg402 of the proteins as the ones which surround the third position of the anticodon (**Figure 3.5c**). Among those residues, the ε-nitrogen of the side chain of Arg402 from *E. coli* GlnRS is observed to establish a polar interaction to the carbonyl group of G36, potentially providing a greater contribution towards the recognition of this position compared to the other more distal amino acids. In the *In*-GlnRS these residues are not completely conserved and align to the amino acids Pro412, Ly413, Tyr414, Lys415, Arg416 (**Figure 3.5c**), respectively. Notably, *In*-GlnRS retains an arginine at the equivalent position to the *E. coli* GlnRS Arg402. I generated a library by randomising these 5 residues of the loop to any of the 20 proteinogenic amino acids, and performed, together with Dr. Shan Tang, a round of positive selection using the dual reporter system composed of *cat*^{112*} and *gfp*^{150*} previously described to identify active variants of the enzyme. Due to the complete lack of activity detectable from the wild type enzyme, the selection was carried out at low concentrations of chloramphenicol on the plate (50 µg/mL). After overnight incubation, GFP-positive clones were identified on the selection plate. The colony showing the highest visible level of fluorescence was isolated and the synthetase variant it harboured, named

In-GlnRS^{S9}, was characterised as the mutant Pro412Asp, Ly413Ser, Tyr414Ala, Lys415His, Arg416Gly,

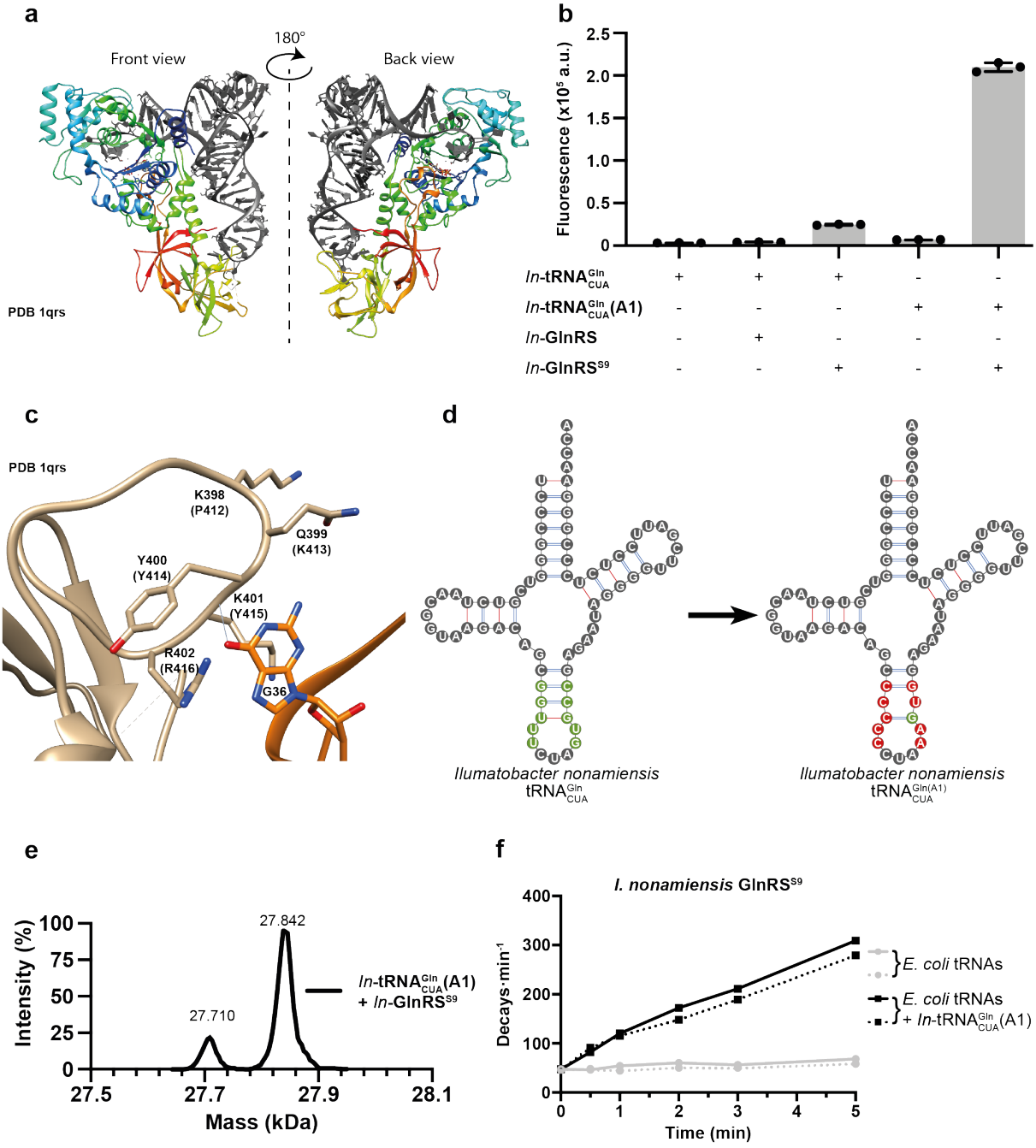


***Figure 3.5**: **a)** Crystal structure of the GlnRS from E. coli (PDB 1qrs[131]) shows how the enzyme folds into a single globular. **b)** Efficiency of the In-tRNA^{Gln}/In-GlnRS pair during different stages of its evolution as an amber suppressor measured using the reporter sfGFP^{150*} revealed how the wild type enzyme did not interact with the amber suppressor mutant of its cognate tRNA. The activity is rescued by the mutations present in the variant In-GlnRS^{S9}. The activity of the pair was further boosted by the engineering of the tRNA anticodon which generated the tRNA^{Gln}(A1). **c)** Inspection of the crystal structure of the E. coli GlnRS allowed the identification of the loop in closest proximity to position 36 of the anticodon which was randomised to optimise the interaction between the synthetase and the amber suppressor variant of the tRNA. **d)** Residues of the anticodon arm (green) randomised and mutation selected (red) in the tRNA^{Gln} to optimise its amber suppression efficiency. **e)** Mass spectrometry of the reporter sfGFP^{150*} confirmed glutamine incorporation by the GlnRS^{S9} mutant. **b)** In vitro aminoacylation by the GlnRS^{S9} mutant highlighted its orthogonality.*

together with the additional mutation Asp558Gly. The new mutant exhibited a small but reproducible amber suppression capacity, as differently from its wild type counterpart it could induce read-through of the sfGFP^{150*} reporter (**Figure 3.5b**).

Overall, this pair behaved similarly to the *Sc*-AspRS, consequently we decided to try optimise the interaction between the synthetase and its substrate tRNA by evolving its anticodon arm, since this strategy had greatly improved the performance of the $Sc-\text{tRNA}^{\text{Asp}}_{\text{CUA}}$ /*Sc*-AspRS pair. We generated a library of tRNA sequences by randomising the 5 nucleotides on either sides of the anticodon of the $In-\text{tRNA}^{\text{Gln}}_{\text{CUA}}$ (**Figure 3.5d**) and repeated the selection to verify which of those variants could be effectively charged by the *In*-GlnRS^{S9}. The selection plate sustained the growth of a variant, which was called $In-\text{tRNA}^{\text{Gln}}_{\text{CUA}}(\text{A1})$, displaying high levels of amber suppression, which was mutated at 9 of the 10 positions randomised and presented an unusual C30:U40 pairing (**Figure 3.5d**). As measured by the production of sfGFP from the amber suppression reporter, the new evolved pair $In-\text{tRNA}^{\text{Gln}}_{\text{CUA}}(\text{A1})$ /*In*-GlnRS^{S9}, was a highly active pair (**Figure 3.5b**), and additionally, mass spectrometry confirmed that the engineering process had not altered the activity of the pair, which was still incorporating glutamine (**Figure 3.5e**).

The importance of the third letter of the anticodon in the recognition of the tRNA by the *In*-GlnRS, highlighted by its complete inactivity of the $In-\text{tRNA}^{\text{Gln}}_{\text{CUA}}$, and the fact that the mutation G36A was unlikely to compromise the interaction between the tRNA and the synthetase due to steric clash, suggested that potentially the enzyme required specific interactions with all the three positions of the anticodon for successful aminoacylation. Consequently, it was possible that the newly evolved mutant *In*-GlnRS^{S9} would recognise specifically the anticodon CUA, which is not present in any of the endogenous *E. coli* tRNAs, hence would have improved orthogonality. I purified the mutant enzyme *In*-GlnRS^{S9} and repeated the *in vitro* aminoacylation assay as I had done previously for the parental wild type enzyme (**Figure 3.5f**). Consistently with the consideration expressed above, the mutant show no detectable incorporation of the labelled amino acid when incubated with the tRNA extract from wild type *E. coli* DH10b, while an increase in aminoacylation was measured over the time course when $In-\text{tRNA}^{\text{Gln}}_{\text{CUA}}(\text{A1})$ was present in the reaction. These results indicated that the evolved pair $In-\text{tRNA}^{\text{Gln}}_{\text{CUA}}(\text{A1})$ /*In*-GlnRS^{S9} is orthogonal in *E. coli*. Notably, the generation of an orthogonal GlnRS/tRNA$^{\text{Gln}}$ amber suppressor pair from the *S. cerevisiae* pair was the first attempt to be made for the genetic code expansion in *E. coli*[132]. Given its high activity, our new pair constitutes a valuable starting point which might be evolved in the future to incorporate ncAAs, even if time constraints prevented me from trying and evolve the pair further.

tRNA^{Glu} from *Sporolactobacillus inulinus*

Glutamyl-tRNA synthetases are monomeric enzymes belonging to the class I of the aaRSs. While in eukaryotes and many prokaryotes these enzymes are exclusively responsible for the aminoacylation of tRNA^{Glu}, in all archaea and several bacterial families they are additionally involved with the synthesis of glutamine, which results from the trans-amination of the side chain of the glutamic acid-bound tRNA^{Gln}. Consequently, the GluRSs from those organisms are non-discriminating between tRNA^{Glu} and tRNA^{Gln}, and for this reason are required to tolerate both C36 found on tRNA^{Glu} and G36 found in tRNA^{Gln}.[133]

The crystal structure of the non-discriminating GluRS from *Thermotoga maritima* (**Figure 3.6a**, PDB 3akz[133]) highlights how the enzyme, in spite of not being clearly subdivided into independent domains connected by loops as in some of the cases observed before, is composed of a catalytic domain at its N-terminus, while the interaction with the anticodons are clustered at the C-terminal domain. In spite of the tolerance that these enzymes must present in order to account for the variations in the anticodon sequences of their substrate tRNAs, conversion of the *Si*-tRNA^{Glu} to an amber suppressor by introduction of the mutations C36A did not result in any appreciable amount of amber suppression (**Figure 3.6b**).

Inspection of the crystal structure of the *T. maritima* GluRS, which was crystallised bound to a tRNA^{Gln}, highlighted how the synthetase did not establish clear contacts with position 36 the anticodon which could account for the lack of aminoacylation of the mutant tRNA (**Figure 3.6c**), however, it allowed me to identify which portion of the enzyme was the most proximal to that residue of the tRNA. In particular, I could identify the loop formed by residues Lys369, Val370, Asn371 and Thr372 in the crystal, which correspond to the residues Tyr370, Gln371, Glu372 and Gln373 in the *Si*-GluRS. Since these residues were not conserved between the two orthologues, it was hard to speculate whether they can establish specific interactions with G36 which would account with the loss of activity on the amber suppressor. I generated a library by randomising those four consecutive residues of the enzyme to any of the 20 proteinogenic amino acids and used the library to perform a step of positive selection using the dual reporter system composed of *cat*^{112*} and *gfp*^{150*} with a low concentration of chloramphenicol (50 μg/mL). After overnight incubation, GFP-positive colonies formed on the selection plates. I isolated the clone which showed the highest amber suppression activity based on visible green fluorescence and verified that it contained a variant,

which was called *Si*-GluRS^{D2}, harbouring the mutations Tyr370Lys, Gln371Gly, Glu372Gly and Gln373Gly, together with the additional spontaneous His375Arg mutation. This mutant displayed a reproducible amber suppression activity (**Figure 3.6b**). Notably, the mutation Tyr370Lys made the

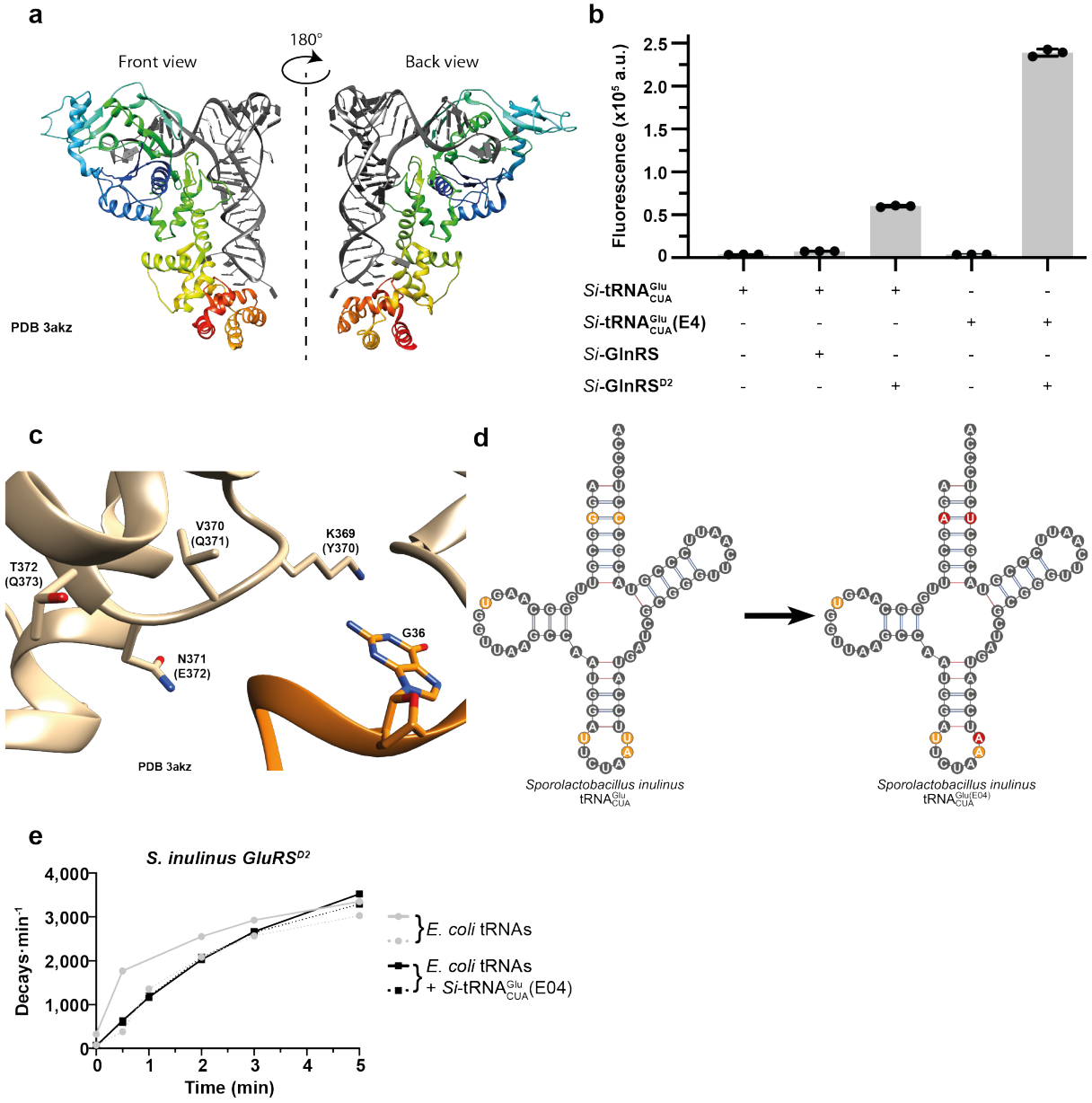


***Figure 3.6**: **a**) Crystal structure of the GluRS from T. maritima (PDB 3akz[133]) highlights how the enzyme is composed of an N-terminal catalytic domain (blue to green) and a C-terminal domain interacting with the anticodon. **b**) Efficiency of the Si-tRNA^{Glu}/Si-GluRS pair during different stages of its evolution as an amber suppressor measured using the reporter sfGFP^{150*} revealed how the wild type enzyme did not interact with the amber suppressor mutant of its cognate tRNA. The tRNA could be recognised by the Si-GluRS^{D2}. The activity of the pair was further boosted by the engineering of the tRNA anticodon which generated the tRNA^{Gln}(E4). **c**) Inspection of the crystal structure of the T. maritima GluRS allowed the identification of the loop in closest proximity to position 36 of the anticodon which was randomised to optimise the interaction between the synthetase and the amber suppressor variant of the tRNA. **d**) Residues of the anticodon arm (yellow) randomised and mutation selected (red) in the tRNA^{Glu} to optimise its amber suppression efficiency. **e**) In vitro aminoacylation by the Si-GluRS^{D2} mutant highlighted its lack of specificity.*

enzyme equal to the crystal structure at this position, however the presence of three consecutive glycines at positions 371-372-373 made it unlikely to believe that the restored activity would be due to the establishment of some specific interactions between the mutated A36 position of the $Si-\text{tRNA}^{\text{Glu}}_{\text{CUA}}$ and the evolved enzyme.

Similar to the strategy implemented for the cases described in the previous sections, I decided to alter the sequence of the tRNA in an attempt to improve the efficiency of the pair, however I only randomised the positions 32, 37 and 38 in the anticodon loop, together with the base pair 3:70 and position 17. A round of positive selection using the dual reporter system allowed me to identify a mutant tRNA which had only acquired the three mutations G3A:C70U and U38A (**Figure 3.6d**). This selected tRNA, which was named $Si-\text{tRNA}^{\text{Glu}}_{\text{CUA}}(\text{E4})$, displayed a significantly increased capacity to suppress the amber stop codon (**Figure 3.6b**).

Given the successful cases in which engineering of the synthetase had lead to an improvement of its orthogonality, I repeated the *in vitro* aminoacylation assay on the *Si*-GluRS^{D2} (**Figure 3.6e**). The data highlighted how the enzyme did not seem to have a faster aminoacylation rate in the presence or absence of the $Si-\text{tRNA}^{\text{Glu}}_{\text{CUA}}(\text{E4})$, similarly to its wild type counterpart (**Figure 2.9**). This seemed to indicate that the specificity of the enzyme had not been improved and that it was still capable of recognising some of the endogenous *E. coli* tRNAs. Since the experimental evidence hinted at the fact that the synthetase would charge multiple tRNAs, we decided that it would not be suitable for genetic code expansion. Considering the high activity of the pair, however, alternative strategies might be considered in the future to further evolve the pair in order to improve its orthogonality, or a similar pipeline could be performed on another pair.

tRNA^{Gly} form *Bacteroides vulgatus* and tRNA^{His} from *Afifella pfennigii*

As discussed previously, the *Bv*-GlyRS could not aminoacylate to any measurable extent the amber mutant of the *Bv*-tRNA^{Gly}. The enzyme is a class II aminoacyl-tRNA synthetase which is active as a homodimer. The structure of the closest orthologue which has been crystallised bound to its tRNA substrate belongs to *Homo sapiens* (**Figure 3.7a**, PDB 4kr2[134]) highlights how the enzyme is composed of an N-terminal catalytic domain and of a C-terminal anticodon-binding domain, similar to other aaRS described in the previous paragraphs. Since the conversion of the tRNA^{Gly} to an amber suppressor involved the mutations C35U and C36A, I identified the residues at the C-terminal domain which are responsible to establish direct interactions with these two positions of the tRNA anticodon. The crystal reveal an extensive network of hydrogen bonds and charge-to-dipole interactions (**Figure 3.7b**). In particular, Thr631 and Gln640 interact with C35, while Arg548 and Arg633 interact with C36. Additionally, Met638 is located in close proximity to C36, position which needed to accommodate a bulkier adenine following the mutation of the anticodon. Consequently, I generated a library of the residues of the *Bv*-GlyRS which correspond to the ones listed above, namely Arg399, Thr478, Arg480, Met485 and Gln487, to evaluate whether a new pattern of interactions could be established between the enzyme and the amber suppressor tRNA which could result in aminoacylation. Unfortunately, no GFP-positive colonies grew on selection plates containing 50 µg/mL of chloramphenicol after a round of positive selection of the library using the dual reporter system.

Alternatively, in order to restore the interaction between the tRNA and the synthetase, Dr. Shan Tang and I tried to generate a library of tRNA sequences where the 5 nucleotides on either sides of the anticodon were randomised (**Figure 3.7c**). Unfortunately, no GFP-positive colonies were selective from this library either after a round of positive selection using the dual reporter system on plates containing 50 µg/mL of chloramphenicol after a round of positive selection of the library.

The histidine-tRNA synthetase is also a class II enzyme which is active as a dimer. The crystal structure of this enzyme from *Thermus thermophilus HB27* (**Figure 3.7d**, PDB 4rdx[135]) highlights how also this enzyme is composed of an N-terminal catalytic domain and of a C-terminal anticodon-binding domain, which are clearly distinct and connected by an unstructured loop (**Figure 3.7d**). Analysis of the contacts between the anticodon and the C-terminal domain of the enzyme highlighted

that G34 of the tRNA contacts the two negatively charged consecutive Asp387 and Glu388 in the crystal, while G36 forms polar contacts with the side chain of Lys397 (**Figure 3.7e**). I generated a library of the equivalent residues of the *Ap*-HisRS, namely Asp483, Glu484 and Lys493 which underwent a round of positive selection using the dual reporter system. Considering that the background aminoacylation of the amber suppressor mutant of *Ap*-tRNA^{His} conferred a resistance up to 250 µg/mL of chloramphenicol, I had to select the library on plates containing 300 µg/mL of antibiotic. Unfortunately, this selection didn't allow the growth of any GFP-positive colony in which

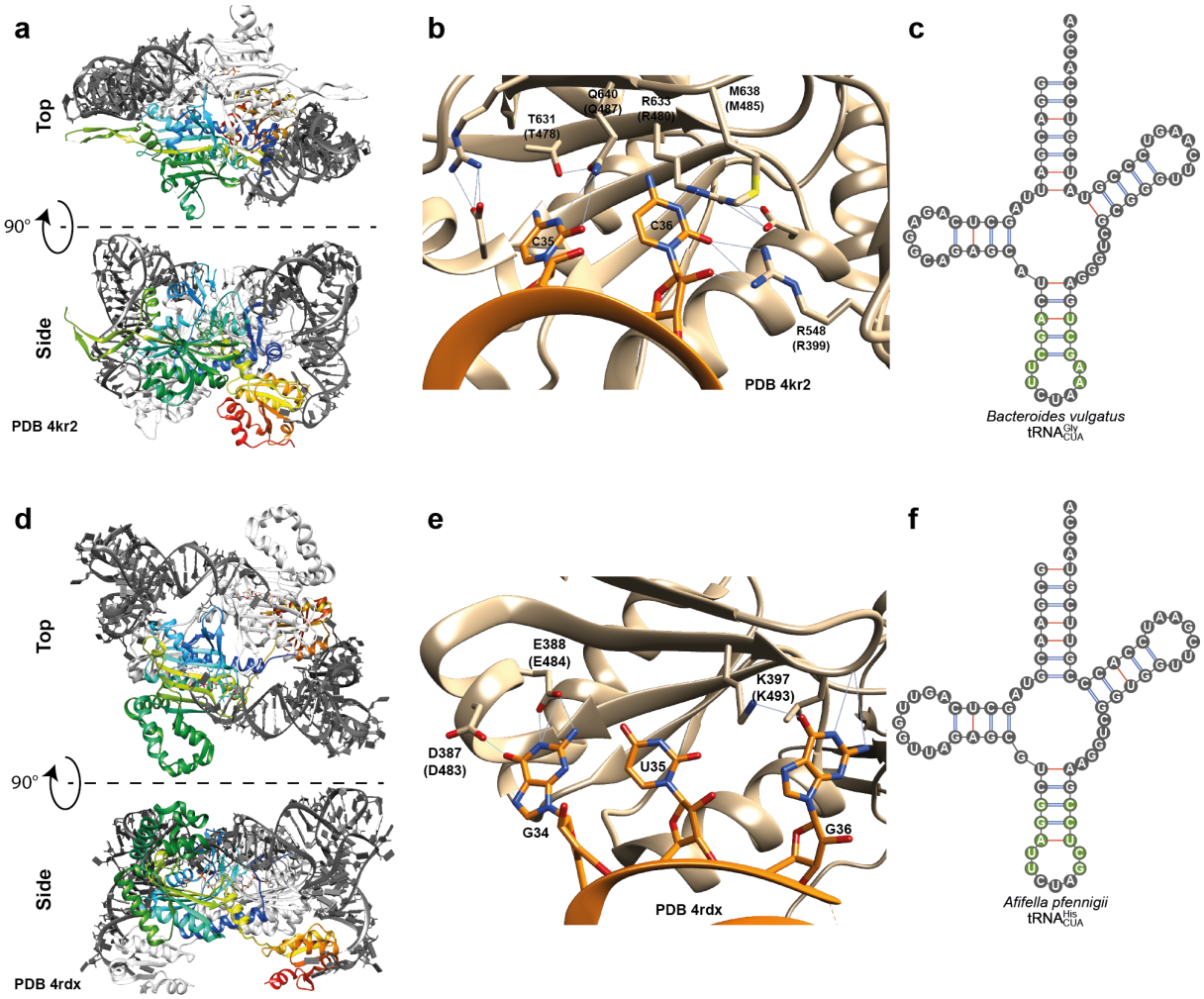


***Figure 3.7**: **a)** Crystal structure of the GlyRS from H. sapiens shows the molecular organisation of the enzyme as a dimer and its division into an N-terminal catalytic domain (blue to green) and a C-terminal anticodon-binding domain (yellow to red). **b)** Inspection of the interaction between the H. sapiens GlyRS and its substrate tRNA reveals a complex network of hydrogen bonds responsible for the specific recognition of the positions C35 and C36. **c)** Positions of the anticodon arm which were randomised in a library (green). No mutants displaying activity were selected.*

***d)** Crystal structure of the His from Thermus thermophilus shows the molecular organisation of the enzyme as a dimer and its clear division into an N-terminal catalytic domain (blue to green) and a distinct C-terminal anticodon-binding domain (yellow to red). **e)** Inspection of the interaction between the T. thermophilus GlyRS and its substrate tRNA reveals hydrogen bonds formed between the positions G34 and G36 of the tRNA anticodon and the C-terminal domain of the enzyme. **f)** Positions of the anticodon arm which were randomised in a library (green). No mutants displaying activity were selected.*

the activity of the enzyme was rescued. In another attempt to obtain an active pair, Dr. Shan Tang and I generated a library of tRNAs in which the 5 nucleotides on either sides of the anticodon were fully randomised which underwent a round of positive selection under the same conditions. Unfortunately also this attempt didn't allow me to identify any mutant pairs in which could perform amber suppression using histidine.

For both of the pairs for which engineering was attempted without success, it remains possible that the failure was connected with an incorrect choice of residues to randomise for the generation of the libraries. It is also possible however that the recognition of the wild type nucleotides in those specific cases cannot be altered to recognised the anticodon with sequence "CUA". In future, these pairs might be successfully evolved to recognise different anticodons which would redirect them towards alternative sense codons made available thanks to the efforts to compress the genetic code of organisms.

tRNA^{Tyr} from *Archaeoglobus fulgidus*

In contrast to all the cases examined in the previous paragraphs, the tRNA^{Tyr} from *Archaeoglobus fulgidus* displayed an unusual behaviour when converted to an amber suppressor by introduction of the single point mutation G34A. In fact, while this tRNA was able to cause read-through of the amber

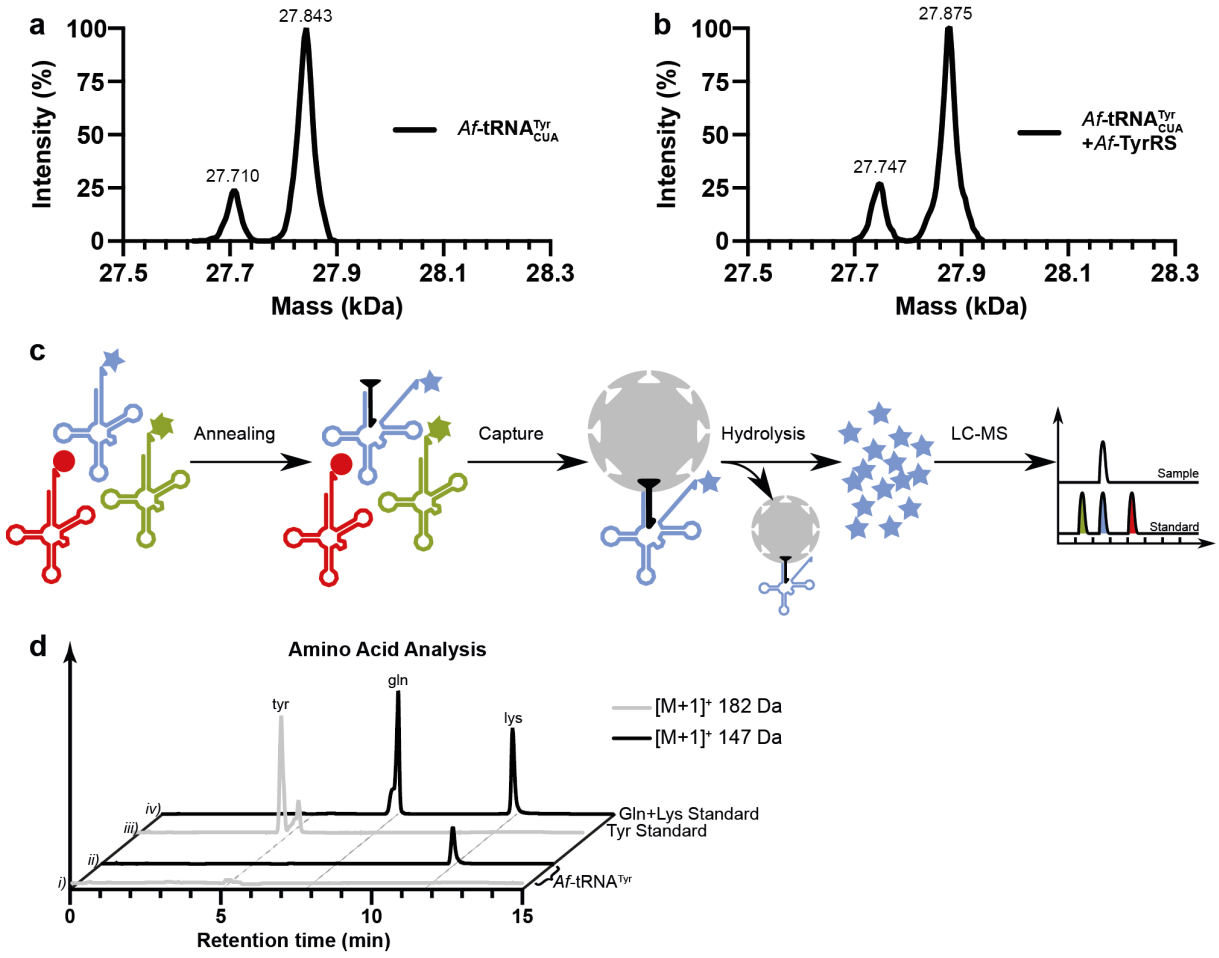


***Figure 3.8**: **a)** Mass-spectrometry analysis of the read-through product of the sfGFP^{150*} reporter induced by the Af-tRNA$^{Tyr}_{CUA}$ highlighted the incorporation of either glutamine, glutamic acid or lysine. **b)** Mass-spectrometry analysis of the read-through product of the sfGFP^{150*} reporter induced by the Af-tRNA$^{Tyr}_{CUA}$ in the presence of its cognate synthetase Af-TyrRS highlighted the incorporation of tyrosine. **c)** Schematic representation of the procedure to identify the amino acid attached to a tRNA of interest. The aminoacylated tRNA sample is isolated and annealed to a tRNA-specific biotinylated DNA probe (Annealing). The probe is then captured onto streptavidin-coated beads (Capture), the unbound tRNAs are washed off, then the amino acid is hydrolysed from the tRNA by alkali treatment (Hydrolysis) and analysed by LC-MS. **d)** Analysis of the amino acid isolated from the Af-tRNA$^{Tyr}_{CUA}$ revealed no signal corresponding to tyrosine (i), while a peak corresponding to the mass and retention time for lysine was observed (ii), as verified by comparison to the relative standards (iii, iv).*

stop codon in the reporter sfGFP^{150*} both in the absence and in the presence of its cognate synthetase, mass spectrometry analysis revealed a qualitative difference in the type of protein produced. In fact, while the purified sfGFP isolated from cells only containing the tRNA had an intact mass of 27.843 Da, which ambiguously identifies either glutamic acid, glutamine or lysine; the mass measured when both the tRNA and the synthetase were present in the cells was equal to 27.875 Da, which uniquely corresponds to the incorporation of tyrosine in response to the stop codon (**Figure 3.8a-b**, **GFP Total Mass**, the minor peaks correspond to removal of Met1 from the translated protein). This evidence suggested that while the synthetase was still capable of recognising the mutant $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$, this tRNA behaved somehow as conditionally orthogonal (i.e. was orthogonal when *Af*-TyrRS was expressed in the cell, but was not orthogonal in its absence). Considering the potential risk of non-specific incorporation of a natural amino acid by the use of a system with these properties, I decided to first focus my attention on the incomplete orthogonality of the tRNA.

In order to minimise the recognition of the $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$ by endogenous *E. coli* aaRSs, I first tried to identify which one among GlnRS, GluRS or LysRS was responsible for the acylation of the amber suppressor. Hence, I developed together with my colleague Dr. Shan Tang an analytical method which allows to characterise which amino acid is bound to a specific tRNA of interest. The experience in developing tREX had showed that extraction of tRNAs at acidic pH allows the aminoacyl ester bond to be preserved. Additionally, tREX had showed that the probe used to detect *Af*-tRNA$^{\text{Tyr}}$, which was designed to not anneal in the proximity of the anticodon, was fully specific (**Figure 2.8** - Tyr_05). We decided to purchase a probe which would contain a biotin instead of the fluorophore Cy5 and to anneal it to a tRNA extract containing $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$ maintaining the acidic pH throughout the procedure. Following annealing, streptavidin-coated sepharose beads were added to the solution and allowed to capture the biotinylated probe together with the aminoacylated $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$ bound to it. The other tRNAs were washed off with a salt buffer at pH 5, then the solution was brought to an alkaline pH to hydrolyse the ester bonds. The amino acids hydrolysed were then collected with the supernatant and analysed by HPLC, while the tRNAs were not eluted from the beads (**Figure 3.8c**).

The procedure was applied to the sample under investigation and the free amino acids were separated on the HILIC-Z column (4.6 x 150 mm, Agilent), and detected using the Agilent 6130 Quadrupole LC-MS unit, ran in SIM mode and set up to monitor either the molecular weight of 182 Da to detect tyrosine, or 147 Da to detect glutamic acid, glutamine or lysine. The chromatographic traces showed that no tyrosine could be detected in solution (**Figure 3.8d** trace *i*), while lysine was the only amino acid detected with a molecular weight of 147 Da (**Figure 3.8d** trace *ii*), as deduced by comparison with the corresponding amino acid standards (**Figure 3.8d** trace *iii* and *iv*).

This result removed the ambiguity in the assignment of the amino acid found in position 150 of the

purified sfGFP and indicated that the *E. coli* LysRS was recognising the $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$ as a substrate. This evidence was somewhat surprising if considering that the mis-acylation was triggered by the he mutation G34C in spite of tRNA^{Lys} having anticodon UUU, but an analogous observation was previously made when the *E. coli* LysRS was shown to be able to weakly mis-charge an amber suppressor tRNA derived from the *S. cerevisiae* tRNA^{Tyr} [136].

In light of this evidence, I decided that I would focus my engineering efforts on two aspects: increase the activity of the pair as an amber suppressor due to specific aminoacylation by the *Af*-TyrRS, and minimise the acylation of the $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$ by LysRS. The *Af*-TyrRS shares 56% of its sequence with the *M. jannaschii* (*Mj*) TyrRS, a synthetase that has been extensively used for genetic code expansion. The crystal structure of *Mj*-TyrRS bound to its tRNA allowed me to identify some features of their interactions (**Figure 3.9a**, PDB 1j1u[137]). In particular, the enzyme is a small protein composed of an N-terminal catalytic domain and a C-terminal anticodon-binding domain. It belongs to the class I of aaRSs but is active as a homodimer. Additionally, recognition of the tRNA anticodon occurs in *trans* within a dimer with respect to the catalytic event, that is, while one monomer binds the anticodon, aminoacylation is carried out by the other monomer. Furthermore, the structure highlights how the contact points between the synthetase and the tRNA are clustered to either the anticodon or the upper part of the acceptor stem.

In order to develop an effective amber suppressor pair, I decided to start off by trying to optimise the interaction between the synthetase and the CUA anticodon, a portion of the tRNA which would not be subsequently modified. In fact, I hoped that the identification of a synthetase with higher activity on the $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$ would be helpful in the event that the evolution of the tRNA to increase its orthogonality would concomitantly result in a reduction of its amber suppression efficiency. This concern was due to the fact that, even if the wild type variant of the enzyme could recognise the $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$, the level of amber suppression was relatively low and comparable to the background caused by the non-specific activity of the *Ec*-LysRS (**Figure 3.9b**).

In order to simplify the selection process for TyrRS variants with higher activity in a context of high background, I decided to adapt the reporter system in use in such a way that a signal would be observed only as a consequence of tyrosine incorporation. I took advantage of the fact that the fluorophore of sfGFP is composed of the triad Thr-Tyr-Gly, in which the aromatic side chain of tyrosine provides a fundamental contribution to the conjugation of the π orbitals required for fluorescence. I hence removed the stop codon at position 150 of the sfGFP^{150*} reporter by mutating it back to its wild type asparagine residue ($\text{sfGFP}^{67\text{Tyr}}$) then I generated a mutant in which the tyrosine in the fluorophore was replaced by lysine ($\text{sfGFP}^{67\text{Lys}}$). As expected, I observed fluorescence from cells expressing $\text{sGFP}^{67\text{Tyr}}$ but not $\text{sfGFP}^{67\text{Lys}}$ (**Figure 3.9c**, black box). Consequently, I tried to generate a new reporter in which the amber stop codon would be replace the tyrosine residue constituting the

fluorophore in position 67 (sfGFP^{67*}). Amber suppression on this reporter induced by the presence of the *Af*− $tRNA^{Tyr}_{CUA}$ together with *Af*-TyrRS resulted in cells displaying visible green fluorescence upon excitation with blue light (~488 nm), which was not observed in the absence of the synthetase (**Figure 3.9c**). This result confirmed that the reporter could be effectively used to monitor amber suppression and it displayed the differential fluorescence required which was not present in its predecessor sfGFP^{150*} (**Figure 3.9c**). Importantly, the sfGFP^{67*} reporter was used from the selections described below, while the old sfGFP^{150*} was used to measure the overall efficiency of amber suppression in a format comparable to previous experiments, which would also allow for quantification of the background.

With the new tool available, I proceeded to identify which residues of the synthetase might be mutated to optimise the interaction between the synthetase and residue C34 of the anticodon (**Figure 3.9d**). Analysis of the crystal structure of the TyrRS from *M. jannaschii* revealed that the wild type G34 is contacted by the side chain of the aspartic acid 286, while the residues Phe261 and Met285 form the lining of a pocket where the nucleotide is accommodated. I hence generated a library of the corresponding residues Phe274, Leu298 and Asp299 of the *Af*-TyrRS by randomising those positions to all the natural amino acid. The library subsequently underwent a round of positive selection on agar plates using the dual reporter system *cat*^{112*} and *gfp*^{67*} and a concentration of chloramphenicol equal to 300 μg/mL. After overnight incubation I identified the most active clones and characterised a mutant, *Af*-TyrRS^{G5}, with the mutations Phe274Val, Leu298Gly and Asp299Arg. Concomitantly, the plasmid harbouring this mutation had gained an extra mutation G63C in the TΨC arm of the tRNA (*Af*− $tRNA^{Tyr}_{CUA}$(G5) , **Figure 3.9e**). This pair displayed an in improved activity and at the same time a reduced background (**Figure 3.9b**). I speculated that the mutation Asp299Arg, which replaces a smaller and hydrogen bond-accepting side chain with a longer, flexible hydrogen bond-donor one, would allow for the establishment of a new compensatory interaction for the mutation G34C. However, as the tRNA concomitantly mutated, I could not unambiguously assign a role to the mutations of the enzyme versus the mutation in the tRNA and I did not perform mutagenesis experiments to confirm my hypothesis.

I then proceeded by engineering the tRNA to obtain a further reduction in the background aminoacylation. I noticed that, differently than the *Af*− $tRNA^{Tyr}_{CUA}$, the *Mj*− $tRNA^{Tyr}_{CUA}$, which has been frequently used in genetic code expansion, did not present a significant background aminoacylation (**Figure 3.9b**). Considering that the identity elements for the *E. coli* LysRS have been poorly characterised, I decided to verify which residues from the *Af*− $tRNA^{Tyr}_{CUA}$ would be identical to the *E. coli* $tRNA^{Lys}$ while being different from the corresponding residues of the *Mj*− $tRNA^{Tyr}_{CUA}$, hoping to identify the positions responsible for the mis-acylation. I hence chose the positions 7:66, 11:20 and 44-45 of the *Af*− $tRNA^{Tyr}_{CUA}$(G5) as targets for randomisation for the generation of a new

tRNA library which underwent a round of positive selection using the dual reporter system described

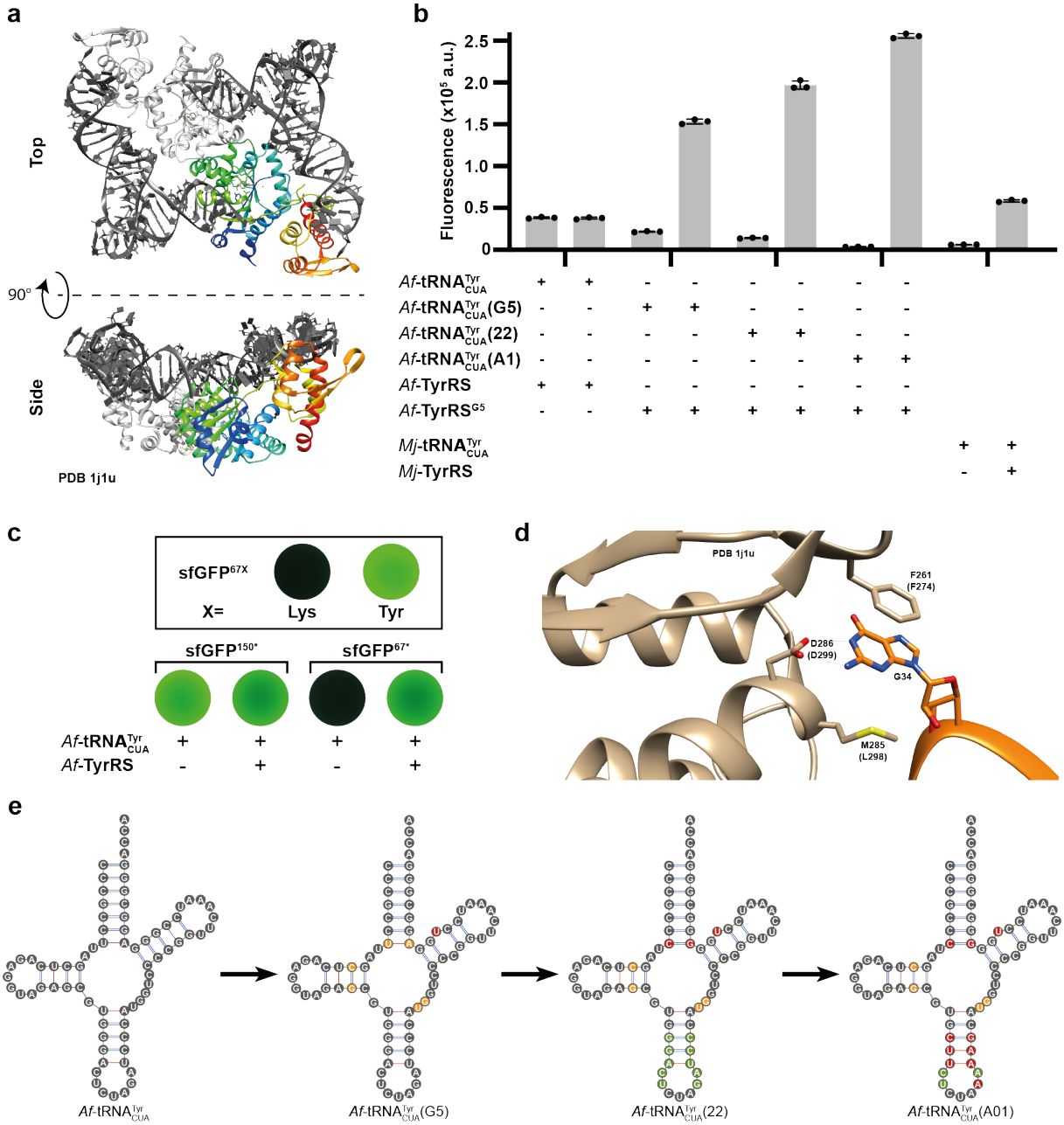


***Figure 3.9**: **a**) Crystal structure of the M. jannaschii TyrRS bound to its tRNA (PDB 1j1u[137]) highlights the interaction in trans between the two domains of the enzyme and the tRNA substrate. **b**) Activity of the A. fulgidus TyrRS/tRNATyr at various stages of its evolution as an amber suppressor, compared to the activity of the amber suppressor pair from M. jannaschii. **c**) E. coli cells expressing a sfGFP variant and a tRNA/aaRS pair as indicated, imaged using a 488 nm laser in a 24 well plate and shown as false-coloured. sfGFP displays fluorescence when position 67 is Tyr, but not when it is Lys (black box). In contrast to sfGFP^{150*}, which can be read through by Af− tRNA$^{Tyr}_{CUA}$ to produces a fluorescence signal both in presence and in absence of the Af-TyrRS, the sfGFP^{67*} becomes fluorescence only in the presence of Af-TyrRS.. **d**) Inspection of the tRNA-binding domain of the Mj-TyrRS allowed the identification of the residues of Mj-TyrRS (equivalent residues of the Af-TyrRS are indicated in parentheses) which surround position G34 in the tRNA. **e**) Evolution of the Af-tRNA^{Tyr} as an amber suppressor. The residues randomised in the first library are shown in yellow, the ones randomised in the second library in green. The selected mutations are shown in red.*

on plates containing chloramphenicol at a concentration higher than before (400 µg/mL)(**Figure 3.9e**). From this selection I isolated a clone where only the base pair U7:A66 was mutated to C7:G66, called $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}(22)$, which displayed an increased signal (**Figure 3.9b**). The background for this tRNA, however, was not completely removed as I would have hoped. Consequently, I decided to generate a new library following a strategy similar to the one used for previous experiments and I fully randomised the 5 residues on either sides of the anticodon of the tRNA, then performed a round of positive selection of plates containing 500 µg/mL of chloramphenicol. By characterising the clones displaying the highest levels of green fluorescence after overnight incubation at 37°C, I identified a mutant tRNA, $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}(\text{A1})$, where 7 out of the 10 randomised positions had been mutated (**Figure 3.9e**), which displayed no appreciable amount of mis-acylation by endogenous *E. coli* synthetases, while having an increased activity as an amber suppressor in combination with its cognate synthetase (**Figure 3.9b**). Interestingly, the highly engineered $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}(\text{A1})$ /*Af*-TyrRS^{G5} displayed ~5-fold higher activity compared to the $Mj-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$ /*Mj*-TyrRS pair.

Engineering the Amino Acid Specificity for *Af*-TyrRS^{G5}

Given that the evolved $Af-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}(\text{A1})$ /*Af*-TyrRS^{G5} resulted highly active and compared favourably with the activity displayed by the well-established $Mj-\text{tRNA}^{\text{Tyr}}_{\text{CUA}}$ /*Mj*-TyrRS pair, I decided to investigate whether it could also be used as an efficient tool for the incorporation of ncAAs in *E. coli*. Interestingly, and differently from all the cases of the synthetases described so far, the crystal structure of the wild type *Af*-TyrRS had been previously resolved (**Figure 3.10a**, PDB 2cyb[138]). This information allowed to confirm that the active site of this enzyme is fully conserved between the orthologues from *M. jannaschii* and *A. fulgidus*. In particular, to allow the specific recognition of tyrosine as a substrate and its discrimination from other amino acids containing aromatic rings (phenylalanine or histidine), the Asp165 residue directly facing the entrance to the active site establishes a charge-to-dipole interaction with the hydroxyl group of the tyrosine side chain, while the Tyr36 residue also contributes to the binding of the substrate by engaging it in a hydrogen bond. Given the similarity between the two orthologues, I decided to transplant the mutations previously characterised for the *Mj*-TyrRS onto the corresponding residues for the *Af*-TyrRS to verify whether this would transfer the capacity to incorporate ncAAs from the *Mj*-TyrRS to the *Af*-TyrRS (**Table 3.3**).

***Mj*-TyrRS residue**	**Tyr32**	**Leu65**	**His70**	**Glu107**	**Asp158**	**Ile159**	**Leu162**
***Af*-TyrRS residue**	**Tyr36**	**Leu69**	**His74**	**Glu114**	**Asp165**	**Ile166**	**Leu169**
*p*BpaRS[139]	Gly			Ser	Thr	Ser	
pN_3PheRS[140]	Thr			Asn	Pro	Leu	Gln
p$B(OH)_2$PheRS[141]	Ser	Ala	Met		Ser		Glu
*p*IPheRS[142]	Leu			Ser	Pro	Leu	Glu
oNO_2TyrRS[143]	His		Cys	Ser	Ser	Ala	Arg

Table 3.3*: Mutations reported in the literature which transform the Mj-TyrRS into the synthetase indicated (left column).*

In spite of the complete identity of the active sites between the two enzymes, both in terms of sequence and 3D geometry, none of the mutant enzymes containing the transplanted mutations displayed any activity. This result was somewhat surprising and difficult to interpret, and it furthermore implied that the previous body of literature on the evolution of the *Mj*-TyrRS could not be directly transposed to my evolved pair. As a consequence, I decided to verify whether new mutations could be selected directly onto the *Af*-TyrRS to alter its amino acid specificity using no further *a priori* knowledge. However, in order to simplify the selection process and prevent the need of repeated cycles of positive and negative selections which were needed in previous studies[140], I decided to verify whether the sfGFP^{67*} reporter, which is effectively a sensor for the incorporation of tyrosine, could be adapted to become a sensor for the incorporation of amino acids other than tyrosine, the major source of background in the evolution of the synthetase.

Given that the fluorescence properties of the fluorophore of the mature GFP are due to the electronic properties of the electron-rich negatively-charged fluorophore[144], and considering that mutants fluorophores display markedly reduced or abolished fluorescence[145], I speculated that the incorporation of most ncAAs within the fluorophore would either interfere with the correct folding of the protein or result in a dysfunctional fluorophore. Under this hypothesis, the sfGFP^{67*} reporter should provide no signal if the desired outcome, i.e.: incorporation of the ncAA, is achieved. In spite of the lack of fluorescence, however, read-through of position 67 would result in the production of a full-length protein. I consequently modified the design of the protein by fusing in frame a short linker followed by a far-red fluorescent protein, E2Crimon146 (**Figure 3.10b**). In this case, I hypothesised that the reporter would behave in one of three possible ways:

1. If tyrosine was incorporated at position 67, a wild-type sfGFP would be produced fused in frame with E2Crimson. Given the spectral separation between the two proteins, the reporter should display both green fluorescence due to sfGFP when irradiated with blue light (~488 nm) and far-red fluorescence when irradiated with red light (~635 nm);
2. If no amber suppression was performed, only the N-terminal portion of sfGFP would be translated and no fluorescence should be observed;

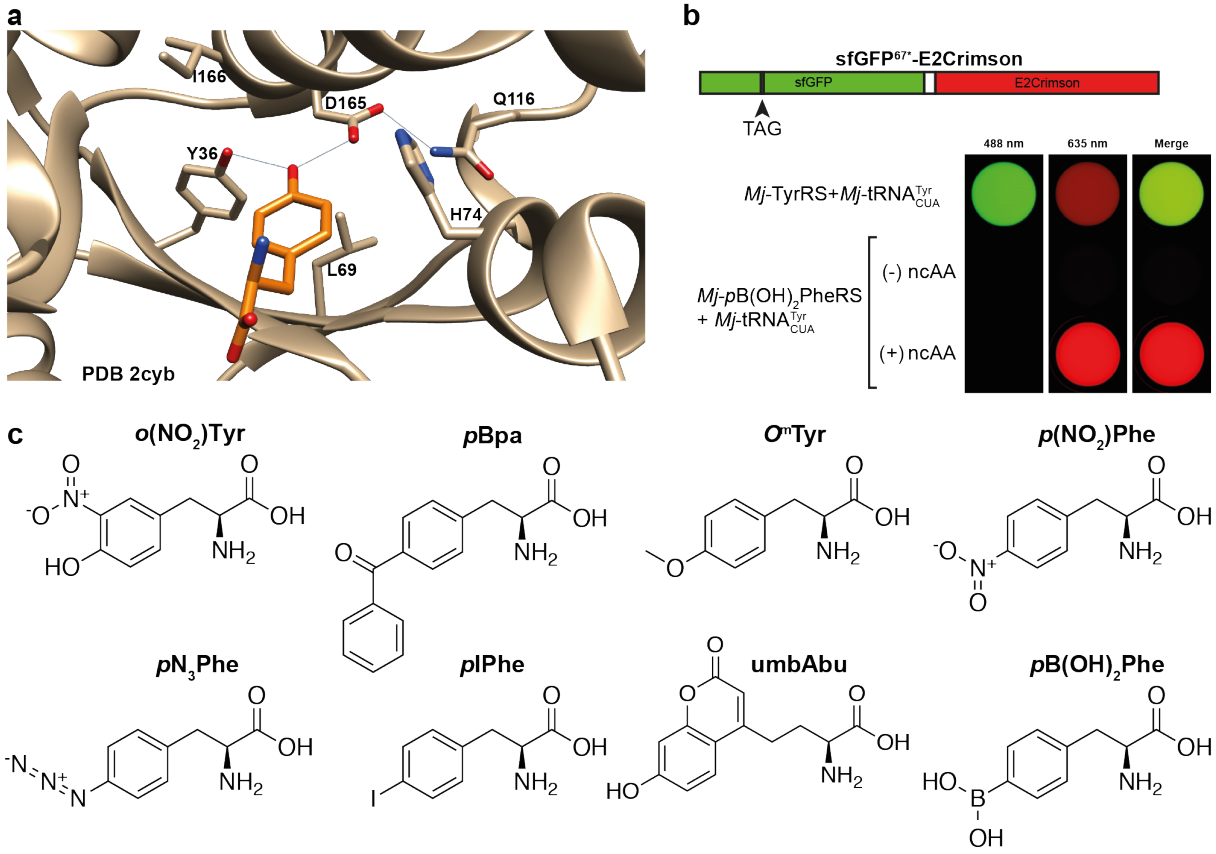

a
I166
D165
Q116
Y36
H74
L69
PDB 2cyb
b
sfGFP67*-E2Crimson
sfGFP
E2Crimson
TAG
488 nm
635 nm
Merge
Mj-TyrRS+Mj-tRNA Tyr CUA
Mj-pB(OH)2PheRS + Mj-tRNA Tyr CUA
(-) ncAA
(+) ncAA
c
o(NO2)Tyr
pBpa
OmTyr
p(NO2)Phe
pN3Phe
pIPhe
umbAbu
pB(OH)2Phe
d
Fluorescence intensity (x10⁵ a.u.)
Af-tRNA Tyr CUA
Af-TyrRS G5
Af-TyrRS H11
D165→ D A N S E D A N S
e
pIPhe
pN3Phe
OmTyr
ncAA
Af-tRNA Tyr CUA
Af-TyrRS mut
Mj-tRNA Tyr CUA
Mj-TyrRS mut
f
Af-tRNACUA(A1)+Af-pIPheRS
27.986
27.853
Mass (kDa)
g
Af-tRNA Tyr CUA(A1)+Af-pN3PheRS
27.898
Intensity (%)
Mass (kDa)
h
Af-tRNA Tyr CUA(A1)+Af-OmTyrRS
27.889
27.757
Intensity (%)
Mass (kDa)

***Figure 3.10**: **a)** Crystal structure of the Af-TyrRS bound to the amino acid substrate (PDB 2cyb[138]) highlights the residues which form the active pocket and the side chains of Tyr36 and Asp165 as the ones directly interacting with the hydroxyl group of the amino acid. **b)** Design of the sensor for the incorporation of ncAAs by evolution of the Af-TyrRS. Fusion of the E2Crimson CDS in frame with the sfGFP[67*] CDS results in a protein displaying green and far-red fluorescence if tyrosine is incorporated in response to the stop codon, but exclusively far-red fluorescence if a ncAA is incorporated instead. **c)** List of the chemical structures of the ncAAs which were used for the selection of mutants of the Af-TyrRS with an altered amino acid specificity. **d)** Activity of the Af-TyrRS mutants G5 or H11 mutant (in which the mutation Tyr36Ile is present) when the residues Asp165 was mutated to Ala, Asn, Ser or Glu highlighted how the presence of a negative charge in that position is both necessary and sufficient for the incorporation of tyrosine. **e)** Activity of the final version of the evolved synthetases pIPheRS, pN$_3$PheRS and O^{m}TyrRS, compared to a corresponding mutant of the Mj-TyrRS described in the literature capable of incorporating the same amino acid. **f)** Mass spectrometry analysis of sfGFP[150*] produced by the $Af\text{-}tRNA^{Tyr}_{CUA}$/Af-pIPheRS confirmed the incorporation of pIPhe. The minor peak corresponds to the cleavage of Met1. **g)** Mass spectrometry analysis of sfGFP[150*] produced by the $Af\text{-}tRNA^{Tyr}_{CUA}$/Af-pN$_3$PheRS confirmed the incorporation of pN$_3$Phe. **h)** Mass spectrometry analysis of sfGFP[150*] produced by the $Af\text{-}tRNA^{Tyr}_{CUA}$/Af-O^{m}TyrRS confirmed the incorporation of O^{m}Tyr. The minor peak corresponds to the cleavage of Met1.*

3. If any amino acid other then tyrosine was incorporated at position 67, the sfGFP would be translated with a compromised fluorophore showing either reduced or impaired fluorescence when irradiated with blue light (~488 nm), however E2Crimson would be produced in its wild-type form and far-red fluorescence should be produced when irradiating it with red light (~635 nm).

In order to confirm that such design would work, I took advantage of the $Mj\text{-}tRNA^{Tyr}_{CUA}$/*Mj*-TyrRS pair, and of its mutant *p*B(OH)$_2$PheRS, capable of incorporating (*S*)-2-amino-3-(4-boronophenyl)propanoic acid (**Figure 3.10c**, *p*B(OH)$_2$Phe). I transformed cells containing the new reporter *gfp*[67*]*e2crimson* with a plasmid containing either the $Mj\text{-}tRNA^{Tyr}_{CUA}$/*Mj*-TyrRS pair or the $Mj\text{-}tRNA^{Tyr}_{CUA}$/*Mj*-*p*B(OH)$_2$PheRS pair (**Figure 3.10b**). After overnight incubation, the cells containing the wild type *Mj*-TyrRS displayed both green and far-red fluorescence, while the ones harbouring the *p*B(OH)$_2$PheRS displayed either no fluorescence in either channels if they had been grown in the absence of the ncAA, or far-red fluorescence only if they had been grown in its presence, confirming that the reporter was working as predicted by its design.

I proceeded by creating a library of the *Af*-TyrRS^{G5} by randomising the residues Tyr36, His74, Gln116, Asp165 and Ile166 due to their proximity to the active site. I then chose the following ncAAs as substrates to perform the selections (**Figure 3.10c**):

1. *ortho*-nitro-tyrosine (***o*(NO$_2$)Tyr**, (*S*)-2-amino-3-(4-hydroxy-3-nitrophenyl)propanoic acid));
2. *para*-benzyl-phenylalanine (***p*Bpa**, (*S*)-2-amino-3-(4-benzoylphenyl)propanoic acid);
3. *O*-methyl-tyrosine (***O*^{m}Tyr**, (*S*)-2-amino-3-(4-methoxyphenyl)propanoic acid);

4. *para*-nitro-phenylalanine (***p*(NO$_2$)Phe**, (*S*)-2-amino-3-(4-nitrophenyl)propanoic acid);
5. *para*-azido-phenylalanine (***p*N$_3$Phe**, (*S*)-2-amino-3-(4-azidophenyl)propanoic acid);
6. *para*-iodo-phenylalanine (***p*IPhe**, (*S*)-2-amino-3-(4-iodophenyl)propanoic acid);
7. umbelliferyl-aminobutirric acid (**umbAbu**, (*S*)-2-amino-4-(7-hydroxy-2-oxochromen-4-yl)butanoic acid);
8. *para*-borono-phenylalanine (***p*B(OH)$_2$Phe**, (*S*)-2-amino-3-(4-boronophenyl)propanoic acid).

Interestingly, sfGFP containing *p*N$_3$Phe in place of Tyr within the chromophore was previously expressed and crystallised[147]. This study confirms that the altered fluorophore is not fluorescent, but highlights that the azido group can be accommodated within the β-barrel. This would suggest that other ncAA with side chains of similar size might be accommodated as well, while it seems unlikely that *p*Bpa and umbAbu would not destabilise the fold. This, however, should not interfere with the functioning of the fusion reporter.

As a preliminary experiment before I performed the selections, I transformed the library into cells containing the dual reporter system *cat*^{112*} and *gfp*^{67*}*e2cimson* in order to verify the abundance within the library of active clones incorporating tyrosine after selection on solid agar containing 50 μg/mL of chloramphenicol. After one overnight growth, the plates were fully covered in a loan of green-fluorescent cells. This overwhelming abundance of mutants with wild type activity explains the need for multiple rounds of selections performed in the past for similar selections on the *Mj*-TyrRS.

In order to understand if this background activity could be reduced to simplify the selections, I tried to verify whether surviving cells harboured the wild type *Af*-TyrRS^{G5} or if specific mutants were present. To my surprise, a large number of the most active clones contained the same mutant synthetase, which I called *Af*-TyrRS^{H11}, with the mutations Tyr36Ile, His74Leu, Gln116Thr. This mutant only retained the amino acid Asp165 among the ones in the closest proximity to the substrate, which seemed to be sufficient for the recognition of tyrosine. Furthermore, the mutation of Tyr36 suggested that, in spite of its direct interaction with the hydroxyl group of the substrate, this residue was dispensable for the activity of the enzyme.

I decided to perform site-directed mutagenesis on the *Af*-TyrRS^{G5} and *Af*-TyrRS^{H11} mutants to test the effect of mutations of residue Asp165 in a context in which Tyr36 was either present or absent (**Figure 3.10d**). I introduced the mutations Asp165Ala-Asn-Ser-Glu to the *Af*-TyrRS^{G5} and the mutations Asp165Ala-Asn-Ser to the *Af*-TyrRS^{H11} and compared their activity as amber suppressors using the sfGFP^{150*} reporter. The fluorescence pattern indicated that loss of the negative charge at position 165 would totally impair the activity of the synthetase both in the presence and in the absence of Tyr36, indicating that Asp165 is both necessary and sufficient to allow incorporation of tyrosine by the

synthetase. The only mutant which displayed an activity was the Asp165Glu mutant, from which I concluded that the charge-to-dipole interaction between the residue 165 and the side chain of the amino acid substrate constitutes the major component of the binding energy between the two and that the reduction in activity of the mutant Asp165Glu was probably due to a suboptimal geometric conformation adopted by the synthetase caused by the presence of an additional carbon in the side chain of glutamic acid.

I consequently decided that, instead of performing multiple round of selections for each of the ncAAs of interest, I would have a better chance of success by generating a better starting library depleted of aspartic and glutamic acid at position 165. To achieve this result, I made use of the script I developed previously and which I had used already to generate a library of the *Sc*-AspRS^{C4}. The script allowed me to identify the combination of degenerate primers with sequence MHG, YRC, KYC, AWC and KGG, to be mixed in a ratio of 6:4:4:2:2 respectively, which would randomise position 165 to all amino acids except for Asp and Glu. I hence generated a new library which underwent a single round of positive selection using the dual reporter system cat112* *gfp*^{67*}*e2cimson*. The plates displayed a combination of colonies with no detectable fluorescence, far-red fluorescence or both green and far-red fluorescence. Among the colonies with far-red fluorescence, I identified the following mutants incorporating the amino acids *p*IPhe, *p*N_3Phe or *O*^{m}Tyr.

***Af*-TyrRS residue**	**Tyr36**	**His74**	**Glu114**	**Asp165**	**Ile166**
*Af-p*IPheRS*	Ile	Leu	Glu	Thr	Gly
Af-pN_3PheRS*	Thr	Leu	Glu	Thr	Gly
Af-O^{m}TyrRS	Ile	Leu	Asn	Phe	His

***Table 3.4**: List of the mutations selected for the Af-TyrRS which allowed the incorporation of pIPhe, pN$_3$Phe and O^{m}Tyr, respectively. The synthetases indicated by * were further engineered.*

The mutants selected displayed an activity comparable to the mutant *Mj*-TyrRSs identified which can incorporate the same amino acids (not shown), but lower than the activity of the wild type pair. In order to verify if their activity could be further boosted, I performed a round of random mutagenesis on the synthetases *Af-p*IPheRS and *Af-p*N_3PheRS and performed a new single round of positive selection and identified the following synthetases:

***Af*-TyrRS residue**	**Tyr36**	**Leu69**	**His74**	**Glu114**	**Asp165**	**Ile166**	**Asn190**
*Af-p*IPheRS	Ile	Met	Leu	Glu	Thr	Gly	
Af-pN_3PheRS	Thr		Leu	Glu	Thr	Gly	Lys

***Table 3.5**: List of the mutations contained in the most active variants of Af-TyrRS engineered to incorporate pIPhe and pN$_3$Phe.*

The new evolved *Af-p*IPheRS and *Af-p*N_3PheRS displayed a significant activity as amber suppressors which was dependent on the presence of the ncAA. Furthermore, they were more active than the

corresponding mutants *Mj*-*p*IPheRS and *Mj*-*p*N_3PheRS (**Figure 3.10e**). The *Af*-*O*^{m}TyrRS displayed comparable activity compared to the *Mj*-*O*^{m}TyrRS but was not further improved (**Figure 3.10e**). Mass spectrometry analysis performed after expression of the sfGFP^{150*} reporter with the *Af*-*p*IPheRS, *Af*-*p*N_3PheRS and *Af*-*O*^{m}TyrRS unambiguously confirmed incorporation of the correct ncAA (**Figure 3.10f-h**, the minor peaks correspond to cleavage of Met1).

This result confirmed that the newly engineered pair can be effectively used for genetic code expansion with an efficiency higher than the one displayed by previously described pairs. Furthermore, the identification of mutants capable of incorporating 3 different ncAA from the same library after a single round of positive selection represented a promising progress which might allow for the fast expansion of the set of ncAA which can be charged by *Af*-TyrRS mutants. Interestingly, the difference in behaviour between the *Af*-TyrRS and the *Mj*-TyrRS might potentially indicate that some amino acids for which no mutant of the *Mj*-TyrRS were identified might be substrates for a *Af*-TyrRS mutant.

Eight Mutually Orthogonal tRNA/aaRS Pairs

The experiments described in the previous sections explain how I attempted to generate efficient amber suppressors from the active pairs which I had previously identified by means of tREX. Among all the pairs under investigation, three were successfully converted into orthogonal and efficient amber suppressors:

1. *Sc*−$tRNA^{Asp}_{CUA}$(10) /*Sc*-AspRS^{C4};
2. *In*− $tRNA^{Gln}_{CUA}$(A1) /*In*-GlnRS^{S9};
3. *Af*− $tRNA^{Tyr}_{CUA}$(A1) /*Af*-TyrRS^{G5}.

Additionally, two pairs were identified as orthogonal in their wild type form:

4. *Ca*− $tRNA^{Arg}_{GCG}$ /*Ca*-ArgRS;
5. *Ap*− $tRNA^{His}_{GUG}$ /*Ap*-HisRS.

I hence decided to verify if these pairs were orthogonal with respect to each other and to three other amber suppressing pairs previously used for genetic code expansion:

6. $tRNA^{Pyl}_{CUA}$ /PylRS from *M. mazei* (*Mm*− $tRNA^{Pyl}_{CUA}$ /*Mm*-PylRS);
7. $tRNA^{Pyl}_{CUA}$(6) /PylRS from *M. alvus* (*Ma*− $tRNA^{Pyl}_{CUA}$(6) /*Ma*-PylRS)[64];

8. $tRNA^{Sep}_{CUA}$(v2.0) /SepRS from *M. maripaludis* (*Mm*− $tRNA^{Sep}_{CUA}$(v2.0) /*Mm*-SepRS)[42].

In order to achieve this goal, starting from a set of plasmids each containing a pair of cognate tRNA/synthetase, I generated a new set of plasmids in which the tRNAs and aaRSs were shuffled to every possible combination of a tRNA and a synthetase. These plasmids were transformed in *E. coli* DH10b cells from which tRNAs were extracted and tested for aminoacylation using tREX (**Figure 3.11**). The gels highlighted that aminoacylation was observed exclusively when the tRNAs under investigation were expressed in combination with their cognate synthetase, indicating mutual orthogonality of all the 8 pairs tested.

In spite of the fact that 6 out of the 8 tRNA tested contained the same anticodon and cannot be used at the same time to encode for different ncAAs, this experiment identified the largest set of exogenous tRNA/synthetase pairs mutually orthogonal in *E. coli* yet. In the future, redirection of one or more of these tRNAs to alternative codons might enable the use of multiple of these pairs together.

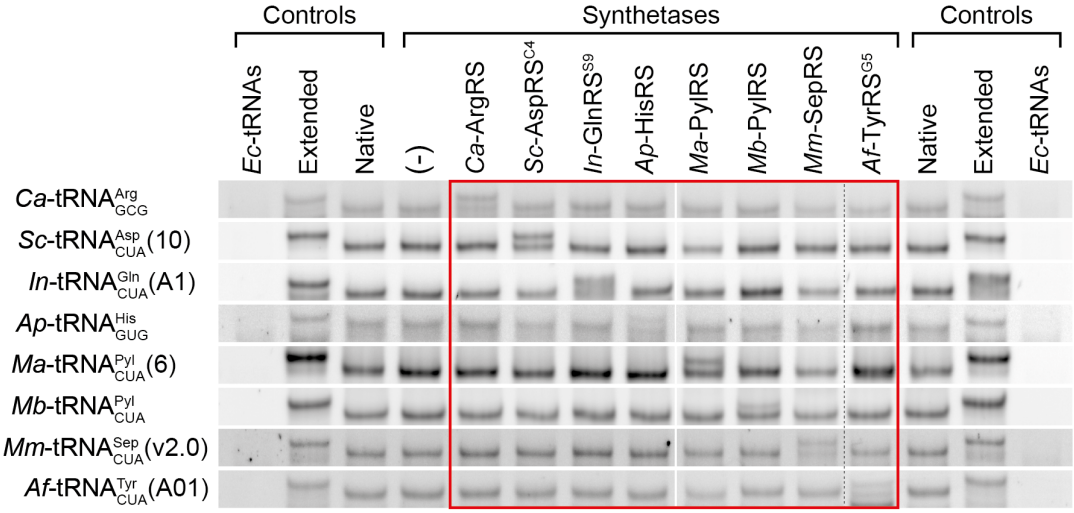


Figure 3.11*: tREX analysis of the cross-amionacylation analysis of the 8 tRNAs listed in combination with each of the 8 corresponding aaRSs showed that none of the non-cognate combinations resulted in aminoacylation of the tRNAs, proving complete cross-orthogonality of these pairs. Samples (-) to Ap-HisRS and Ma-PylRS to Af-TyrRS for each tRNA had to be run on different gels. For this reason, a set of controls is present on each gel. The dashed line between the lanes corresponding to Mm-SepRS and Af-TyrRS corresponds to an empty lane which was removed.*

Discussion

In this chapter I have described my attempts to convert some of the active pairs identified in the screening I illustrated in the previous chapter into amber suppressors. This conversion was needed primarily due to the fact that, in spite of the recent attempts to alter the genetic code of living organisms to free up codons and allow genetic code expansion[30, 148], the research in this field was still at an early stage, so that amber suppression still represented the most common method used to expand the genetic code. Importantly, though, re-direction of a specific tRNA to a different codon constitutes a major alteration in its function and often it abolishes its interaction with its cognate aaRS, while also occasionally inducing mis-charging by other of the endogenous synthetases.

Initial experiments showed that direct conversion of the tRNA under investigation to an amber suppressor caused both loss of orthogonality and loss of recognition by the parent synthetase. However, since the parental wild type variant of those pairs had been investigated and characterised as functional, these changes to the functioning of the pairs was directly imputable to the direct alteration of the anticodon. These occurrences should also serve as an explanation to the choice which was made originally to investigate the pairs in their wild type form before attempting any engineering on them.

I firstly focused my attempts to the evolution of the corresponding synthetases in order to rescue, or optimise their capacity to aminoacylate the amber suppressing mutant of their cognate tRNAs. I consequently managed to obtain active variants for the *Sc*-$tRNA^{Asp}_{CUA}$/*Sc*-AspRS, *Mp*-$tRNA^{Cys}_{CUA}$/*Mp*-CysRS, *In*-$tRNA^{Gln}_{CUA}$/*In*-GlnRS, *Si*-$tRNA^{Glu}_{CUA}$/*Si*-GluRS and *Af*-$tRNA^{Tyr}_{CUA}$/*Af*-TyrRS. Unfortunately, *Ca*-$tRNA^{Arg}_{CUA}$ seemed to be a substrate of the endogenous *E. coli* ArgRS, which prevented me from attempting to engineer the pair further. Conversely, the selections to engineer the *Bv*-GlyRS and the *Ap*-HisRS to accept the *Bv*-$tRNA^{Gly}_{CUA}$ and the *Ap*-$tRNA^{His}_{CUA}$ did not give positive outcomes. In these two cases, though, I could not tell whether this result was due to the intrinsic incapacity of any mutant of those synthetases to recognise the CUA anticodon, or whether more extensive mutations would allow to obtain the desired activity. Regardless, these pairs might be used for the re-assignment of alternative sense codons which will be made available in the future thanks to genome engineering, as discussed earlier.

The five pairs *Sc*-$tRNA^{Asp}_{CUA}$/*Sc*-AspRS, *Mp*-$tRNA^{Cys}_{CUA}$/*Mp*-CysRS, *In*-$tRNA^{Gln}_{CUA}$/*In*-GlnRS,

Si− $tRNA^{Glu}_{CUA}$ /*Si*-GluRS and *Af*− $tRNA^{Tyr}_{CUA}$ /*Af*-TyrRS were additionally optimised to maximise their performances as amber suppressors. Importantly, as a consequence of these alterations the engineered variants of the two synthetases *Sc*-AspRS and *In*-GlnRS, named *Sc*-AspRS^{C4} and *In*-GlnRS^{S9} respectively, showed a marked improvement to their orthogonality with respect to the endogenous *E. coli* tRNAs. On the other hand, *Si*-GluRS^{D2} displayed the same lack of discrimination as its parental counterpart, while neither *Mp*-CysRS nor its active truncation product could be tested.

In addition to optimisations of the pairs as amber suppressors, extensive investigations were performed on the possibility to alter the amino acid specificity for several of the synthetases indicated. Among those, engineering of the *Sc*-AspRS^{C4} made possible by a clever design of the library of mutants allowed the identification of a variant capable of incorporating glutamic acid instead of aspartic acid. Even if the difference between these two amino acids is small and even if glutamic acid is a natural proteinogenic amino acid, this result was a proof of principle indicating that the synthetase can be modified to alter its amino acid recognition. Future attempts will be performed to incorporate ncAAs using this system.

Another important achievement which resulted from the work described was the evolution of three ncAA-RS redived from *Af*-TyrRS^{G5}. In spite of its similarity with the well-characterised *Mj*-TyrRS and the complete three-dimensional conservation of the residues forming its active pocket, this enzyme behaved differently from the *M. jannaschii* orthologue as none of the attempts of transplanting mutations described on the *Mj*-TyrRS onto the *Af*-TyrRS^{G5} resulted in active mutants. As a consequence, a *de novo* search for variants capable of incorporating ncAAs was performed, taking advantage of an improved design for the library generation and of a dedicated reporter system capable of discriminating the incorporation of a ncAA from the incorporation of tyrosine, the wild type substrate for the synthetase. As a result, mutants were found from the same library which could incorporate three distinct amino acids, namely *p*-iodo-L-phenylalanine, *p*-azido-L-phenylalanine, and *O*-methyl-L-tyrosine. The first two were further optimised to achieve incorporation efficiencies significantly higher than the ones achieved using the *Mj*-TyrRS mutant described in the literature capable of incorporating the same amino acids. Importantly, in all likelihood the pool of ncAAs which can be incorporated using the *Af*-TyrRS^{G5} will be expanded in the future by screening of libraries with alternative designs.

Finally, I verified that the three evolved amber suppressor pairs *Sc*− $tRNA^{Asp}_{CUA}$ /*Sc*-AspRS^{C4}, *In*− $tRNA^{Gln}_{CUA}$ /*In*-GlnRS^{S9}, *Af*− $tRNA^{Tyr}_{CUA}$ /*Af*-TyrRS^{G5}, the two native orthogonal pairs *Ca*− $tRNA^{Arg}_{GCG}$ /*Ca*-ArgRS and *Ap*− $tRNA^{His}_{GUG}$ /*Ap*-HisRS and the three previously described amber suppressor pairs *Mm*− $tRNA^{Pyl}_{CUA}$ /*Mm*-PylRS, *Ma*− $tRNA^{Pyl}_{CUA}$(6) /*Ma*-PylRS and *Mm*− $tRNA^{Sep}_{CUA}$(v2.0) /*Mm*-SepRS display no cross-reactivity to one another and could in principle

be used at the same time within a cell. While constituting the largest set of mutually orthogonal exogenous pairs of tRNA/synthetase in *E. coli* described so far, the actual concomitant use of multiple of these pairs within a cell will depend on future efforts to re-direct the tRNAs to distinct codons.

Overall, during my project I managed to establish a procedure which enabled us to experimentally test the behaviour of exogenous aaRS/tRNA pairs transplanted into *E. coli*. This pipeline allowed us to identify several new active pairs, some of which were further engineered as amber suppressors. The evolution of aminoacyl-tRNA synthetases to incorporate ncAA was successful for the tyrosyl-tRNA synthetase from *A. fulgidus,* which could be mutated to accept amino acids containing mostly hydrophobic side chains with an efficiency comparable or higher than previously described systems. This result confirmed that engineering specific hydrophilic interaction is significantly more challenging, while more hydrophobic active sites display more flexibility in their tolerance for substrates. The methodology developed will allow future identification of orthogonal pairs for any of the desired isoacceptor classes.

Chapter IV – Materials & Methods

The following sections are adapted from my publication, "Cervettini, D., Tang, S., Fried, S.D. *et al*. Rapid discovery and evolution of orthogonal aminoacyl-tRNA synthetase–tRNA pairs. *Nat Biotechnol* **38,** 989–999 (2020)."

Materials

Arabinose, antibiotics, IPTG, liquefied phenol, chloroform, *p*-iodo-L-phenylalanine (CAS 24250-85-9, cat. No. I8757-1G), and *O*-methyl-L-tyrosine (CAS 6230-11-1, cat. No. 158259-1G) were purchased from Sigma-Aldrich. *p*-azido-L-phenylalanine (CAS 33173-53-4, cat. No. 4020250.0005) was purchased from Bachem. All ^{14}C-labelled radiochemicals (arginine: cat. No. MC137; aspartic acid: cat. No. MC139; glutamic acid: cat. No. MC 156; glutamine: at. No MC1124; glycine: cat. No. MC163; isoleucine: cat. No. MC174; proline: cat. No. MC263; tyrosine: cat. No. MC275) and ^{3}H-labelled histidine (cat. No. MT905) were purchased from Moravek Inc.

tRNA Alignment

tRNA sequences were downloaded from tRNA-DB-CE together with the information about the secondary structure of the candidate tRNAs, which constituted the basis for the alignment performed (http://trna.ie.niigata-u.ac.jp/cgi-bin/trnadb/index.cgi). The sequences were sorted based on their isoacceptor classes. The database contains the following structural information:

1→7	Acceptor stem, first strand	**39→43**	Anticodon stem, second strand
8→9	Unpaired bases	**44→48**	Variable loop
10→13	D arm, first strand	**49→53**	TΨC stem, first strand
14→21	D loop	**54→60**	TΨC loop
22→25	D arm, second strand	**61→65**	TΨC stem, second strand
26	Unpaired base	**66→72**	Acceptor stem, second strand
27→31	Anticodon stem, first strand	**73**	Discriminator base
32→38	Anticodon loop		

To convert this information into an alignment of tRNAs composed of canonical positions 1 to 76, the following procedure was performed:

i) for tRNAs with an 8-nucleotide acceptor stem, the first 7 base pairs were assigned to positions 1→7/66→72. The only *E. coli* tRNA with such feature, $tRNA^{SeC}$, aligns in such manner to the $tRNA^{Ser}$s, with which it shares the recognition by the seryl-tRNA synthetase;

ii) if the D arm is formed of 3 base pairs, positions 13 and 22 are assigned to the first and last nucleotides of the D loop;

iii) the D loop presents a very high degree of variability, and commonly lacks the two distinctive Gs at positions 18 and 19 found in *E. coli* tRNAs. For these positions, the alignment was manually generated as shown in **Chapter V – Appendix**: **D Loop Alignment Table**;

iv) as the variable loop can have a very broad range of lengths, the positions 44→48 were assigned in the following order: 44→48→45→47→46. For loops longer than 5 nucleotides, the first two positions are numbered 44 and 45, while the last three are numbered 46→48;

v) positions 74→76 always have sequence CCA, either because the tRNA gene natively has such sequence, or because the tRNA processing enzymes edit the original sequence and *a posteriori* add it.

The sequence alignment of *E. coli* tRNAs is shown in **Figure 2.1**.

Computational Analysis

Candidate tRNAs, aligned as described above, were compared individually to each of the distinct 47 *E. coli* isoacceptor tRNAs, which were also aligned as described above (see **Figure 2.1**). For each candidate tRNA ($tRNA^{X}$) and *E. coli* isoacceptor (e.g.: *E. coli* $tRNA^{Ala}_{GGC}$), the nucleotides of the candidate $tRNA^{X}$ at the positions of the identity elements for the *E. coli* isoacceptor (e.g: positions 2, 3, 4, 20, 69, 70, 71, 73 for *E. coli* alanine isoacceptors) are scored. Each position of $tRNA^{X}$ is scored +1 if it matches the nucleotide at that position of that *E. coli* (e.g.: *E. coli* $tRNA^{Ala}_{GGC}$) isoacceptor, otherwise it is scored -1. The final score for this pairwise comparison (e.g: $tRNA^{X}$ to *E. coli* $tRNA^{Ala}_{GGC}$) is the average of the individual scores across the positions of the identity elements. The score of $tRNA^{X}$ for *E. coli* AlaRS is the average score of $tRNA^{X}$ for each of the *E. coli* $tRNA^{Ala}$ isoacceptors (e.g: *E. coli* $tRNA^{Ala}_{GGC}$ and $tRNA^{Ala}_{TGC}$). The complete list of identity elements for *E. coli* aaRSs used in this work is shown in the next section**.** Following the scoring procedure, tRNAs with a score for the same *E. coli* aaRS as their isoacceptor class greater than 0.0 were filtered out. A subset of the remaining tRNAs were chosen for experimental characterisation (**Chapter V – Appendix: tRNA Experimentally Tested**). tRNAs for selenocysteine and natural suppressor tRNAs were not considered for experimental characterisation.

Table of Identity Elements

Below is the list of all the identity elements used from the scoring algorithm in this work, derived from analysis of the literature[20, 118-120].

Identity elements for *E. coli* aaRSs

AlaRS	2, 3, 4, 20, 69, 70, 71, 73
ArgRS	20, 35, 36, 38, 73
AsnRS	34, 35, 36, 73
AspRS	2, 10, 34, 35, 36, 38, 71, 73
CysRS	2, 3, 13, 15, 22, 34, 35, 36, 48, 70, 71, 73
GlnRS	1, 2, 3, 10, 34, 35, 36, 37, 38, 70, 71, 72, 73

GluRS	1, 2, 11, 13, 22, 24, 34, 35, 37, 46, 71, 72
GlyRS	1, 2, 3, 35, 36, 70, 71, 72, 73
HisRS	4, 34, 35, 36, 69, 73
IleRS	4, 12, 23, 29, 34, 35, 36, 37, 38, 41, 69, 73
LeuRS	14, 15, 16, 47, 48, 73
LysRS	34, 35, 36, 73
MetRS	4, 5, 34, 35, 36, 68, 69, 73
PheRS	20, 27, 28, 34, 35, 36, 42, 43, 44, 45, 59, 60, 73
ProRS	15, 35, 36, 48, 72, 73
SerRS	2, 3, 4, 11, 24, 69, 70, 71, 72, 73
ThrRS	1, 2, 35, 36, 71, 72
TrpRS	1, 2, 3, 34, 35, 36, 70, 71, 72, 73
TyrRS	1, 34, 35, 46, 72, 73
ValRS	3, 4, 35, 36, 69, 70, 73

Plasmid Generation

Our standard expression plasmid for tRNA/synthetase pairs consists of the pBR322 origin of replication (high copy number); the constitutive *E. coli glnS* promoter which controls the expression of the aminoacyl-tRNA synthetase of interest followed by a spacer; the strong constitutive *E. coli lpp* promoter which controls expression of the tRNA of interest and an *rrnC* terminator following the tRNA. All plasmids were constructed starting from the plasmid containing PylRS and $tRNA^{Pyl}$ from *M. mazei* and removing the coding sequence of the synthetase from the start ATG codon to the last codon of the CDS (the TAA stop codon was left in place) by NEB HiFi DNA assembly (Cat. No. E2621L). The remaining plasmid containing only a tRNA was used as a template in which $tRNA^{Pyl}$ was systematically replaced by the tRNA under investigation ($tRNA^{X}$). These plasmids are called pKW-($tRNA^{X}$) (e.g. pKW-*Af*-$tRNA^{Tyr}$). For those tRNAs which were investigated together with their cognate synthetase, this synthetase was re-introduced where the PylRS CDS was located. These plasmids are called pRST-$tRNA^{X}$-aaRS (e.g. pRST-*Af*-$tRNA^{Tyr}$-*Af*-TyrRS). All tRNAs libraries were generated on pKW plasmids and all synthetases libraries were generated on pRST plasmids.

The reporter plasmid p15A-cat^{112*}-gfp^{150*} contains the p15A origin or replication, a constitutively expressed tetracycline resistance gene, in addition to the constitutive positive selection marker cat^{112*}

(chloramphenicol acetyl transferase gene interrupted by an amber stop codon at position 112) and the fluorescent reporter *gfp*^{150*} (the sfGFP gene containing the amber stop codon at position 150) under the control of the L-arabinose-inducible P_{BAD} promoter. This plasmid is used for selections, unless otherwise stated.

The p15A-*cat*^{112*}-*gfp*^{67*} reporter plasmid was derived from the p15A-*cat*^{112*}-*gfp*^{150*} plasmid by mutating codon 150 to AAG (Asn codon) and codon 67 to TAG.

The p15A-*cat*^{112*}-*gfp*^{67*}*e2crimson* plasmid was derived from the p15A-*cat*^{112*}-*gfp*^{67*} plasmid by fusing the *gfp* gene in frame with a linker and the *e2crimson* gene.

tRNA Extraction

The tRNA extraction described below was adapted from a previous method[149]. A pre-culture of the desired strain was incubated overnight at 37°C with shaking at 220 rpm. 2 ml of pre-culture was diluted into 50 ml of pre-warmed rich medium (2xYT) and the cells were incubated at 37°C with shaking at 220 rpm for 1-2h. At OD_{600}=0.5~1.0, cells were pelleted (5 min at 5,000 rcf) and the supernatant removed. The cell pellet was resuspended in 800 µl of buffer D (composition below) and transferred to a 2 ml tube. Cells were pelleted again (2 min at 5,000 rcf), the supernatant discarded and the pellet was resuspended in 450 µl of buffer D. 50 µl of liquefied phenol (90% phenol in water from Sigma Aldrich, cat. No. P9346) was added to the cells as a lysing agent, and the lysis was performed by head-over-tail rotation (15 rpm, 15 min, room temperature). The lysed cells were pelleted (25 min, >20,000 rcf, 4°C) and ~500 µl of supernatant containing the tRNAs was recovered and transferred to a clean tube. 500 µl of chloroform was added to the solution and the tube was thoroughly vortexed for 1 min until a cloudy emulsion was formed. The emulsion was separated by centrifugation (1 min at >20,000 rcf). The top layer containing the tRNAs was recovered (~480 µl). This solution was either frozen at -20°C or immediately processed as described in **tREX Protocol**.

Buffer D composition

NaOAc 50 mM pH 5	NaCl 150 mM	$MgCl_2$ 10 mM	EDTA 0.1 mM

tREX Probes Design

The tREX probes are DNA probes composed of two sections. The 3′-section of the probes is the reverse complement of the sequence of the tRNA of interest between canonical position 45 (e.g: from the second nucleotide of the variable loop) to canonical position 76 (e.g: to the 3′-CCA end of the tRNA). The 5′-section of the probe is a poly-A sequence, whose length scales with the length of the 3′-section as follows:

Length of the 3′-section	30 to 31	32 to 33	34 to 36	37 to 38	39 to 41	42 to 43	44 to 46	47 to 48	49 to 51
Number of As	12	13	14	15	16	17	18	19	20

Probes were purchased from Sigma Aldrich as PAGE-purified oligos and were labelled at the 3′-OH with the cyanine 5 (Cy5) dye. The sequence of all the probes used for the screening is reported in **Chapter V – Appendix: tRNA Experimentally Tested**.

tREX Protocol

The tRNA extract was aliquoted into 3x136 µl samples (A, B and C).

Sample A (positive control for full extension) was brought to 160 µl with buffer D (composition below) and the tRNAs were precipitated with 375 µl of absolute ethanol. After incubation (1h, 10°C) the sample was centrifuged (25 min at >16,100 rcf), the supernatant removed and the pellet dried. The pellet was then resuspended in buffer D to a final concentration of 1 µg/µl, as measured by nanodrop.

Sample B (negative control for lack of extension) was deacylated by adding 8 µl of NaOH 300 mM (42°C, 1h), then neutralized by addition of 8 µl NaOAc 3 M, pH 5, and oxidized by addition of 8 µl of $NaIO_4$ 100 mM (1h, 10°C). Next, 375 µl of absolute ethanol was added and the tRNA was precipitated and resuspended as for sample A.

Sample C was brought to 152 µl with buffer D, then oxidized by addition of 8 µl of $NaIO_4$ 100 mM (1h, 10°C). Next, the tRNAs were ethanol precipitated with 375 µl of absolute ethanol and resuspended as sample A and B.

The enzymatic reactions were assembled as follows:

tRNA (1 µg/µl)	dNTPs (10 mM each)	NEBuffer 2.1	Cy5-labelled DNA probe	ddH_2O
2 µl	1 µl	4.5 µl	1 µl	36.5 µl

One reaction was assembled using each of the tRNAs from samples A, B, C, or a control for probe specificity, which consisted of a tRNA extract from wild type *E. coli* DH10b treated like sample C. The reactions were annealed in a thermocycler using the following settings:

95°C	→	70°C	→	50°C	→	4°C
1 min		2 min		2 min		∞

After annealing, the following mix containing the Klenow Fragment (3'→5' exo-) of *E. coli* DNA polymerase I (NEB M0212S) was added to each reaction.

Klenow exo$^-$	NEBuffer 2.1	ddH_2O
0.5 µl	0.5 µl	4 µl

The samples were incubated at 37°C for 20 min.

Next, each reaction was mixed with 50 µl of loading dye 2x (8 M urea, 0.04% Orange G) and 10 µl was run on an acrylamide gel (200 V, approximately 45 min).

Acrylamide 19:1	TBE	TEMED	$(NH_4)_2S_2O_8$
8%	1x	0.1%	0.1%

Gels were imaged on the RGB Typhoon Imager using the red laser (635 nm).

Northern Blot of Aminoacylated tRNAs

tRNA extracts containing the tRNA of interest (see **Chapter IV – Materials & Methods: tRNA Extraction**) were ethanol-precipitated by adding 7/3 volumes of absolute ethanol (1h, 10°C), then resuspended in buffer D (composition below) to a final concentration of 1 µg/µl, as assessed by nanodrop. 2 µg of the tRNA sample was run for 4.5h at 4°C on a 15 cm long acidic urea PAGE (composition below) using 300 mM NaOAc, pH 5, as running buffer and 12 W constant power (~140 V, ~85 mA). The composition of the loading dye (2x) used is provided below.

Following electrophoresis, the gel was stained with SYBR Gold to identify the position of tRNAs. An appropriate section of the gel was cut out and transferred onto nylon membrane (Ambion®

BrightStar®-Plus Positively Charged Nylon Membrane) using the iBlot™ DNA Transfer Stack for iBlot® Dry Blotting System (the membrane contained in the transfer stack is replaced). The tRNAs were cross-linked onto the membrane (Stratalinker® UV Crosslinker 2400), which was later immersed for 20 min in Ambion® ULTRAhyb®-Oligo buffer for blocking. The biotinylated DNA probe with sequence 5′-TGGCGGAAACCCCGGGAATCTAACC CGGCT-3′ to detect $tRNA^{Pyl}$ was then added to a final concentration of 3 ng/µl and hybridised overnight at 37°C. Excess probe was washed off using 2xSSC buffer and then blotting was performed using the Thermo Scientific Pierce Chemiluminescent Nucleic Acid Detection Module.

Buffer D composition

NaOAc 50 mM pH 5	NaCl 150 mM	$MgCl_2$ 10 mM	EDTA 0.1 mM

Acidic urea PAGE

6.5% acrylamide 19:1	100 mM NaOAc pH 5	8 M urea	0.1% TEMED	0.1% $(NH_4)_2S_2O_8$

Loading dye (2x)

100 mM NaOAc pH 5	8 M urea	0.1% xylene cyanol FF	0.1% bromophenol blue

Synthetase Purification

The synthetases of interest were cloned in a pET expression vector under the control of a T7 promoter. To allow for a one-step affinity purification, a StrepTag followed by the TEV cleavage site were both added after the initial methionine (e.g: MWSHPQFEKGS*ENLYFQG*[...]) by using the nucleotide sequence (the StrepTag sequence is underlined, the TEV cleavage side in italics.):

```
ATG TGG AGC CAC CCG CAG TTC GAA AAA GGG AGT GAA AAC CTG TAC TTC CAA GGT […]
Met Trp Ser Bis Pro Gln Phe Glu Lys Gly Ser Glu Asn Leu Tyr Phe Gln Gly […]
```

The synthetase expression plasmids were transformed in *E. coli* BL21 (DE3) cells. 1 L of cell culture was grown in 2xYT after overnight expression at 20°C with shaking at 220 rpm and induction at OD_{600} = 0.5 with 0.2 mM IPTG. Cells were harvested at 5,000 rcf for 10 min, resuspended in 2 volumes of wash buffer containing cOmplete™ Protease Inhibitor Cocktail (Roche) and sonicated. The debris were sedimented at 18,000 rcf and the supernatant was mixed with 1 mg of avidin and loaded onto a StrepTrap HP column (GE) for affinity purification of synthetase protein. The column was washed with Wash buffer until the UV trace at 280 nm of the measured flow-through returned to baseline,

then the column was equilibrated in Elution Buffer minus desthiobiotin before the protein of interest was eluted with Elution Buffer and concentrated. The composition of the buffers was as follows:

Wash Buffer	Elution Buffer
50 mM sodium phosphate buffer pH 8	20 mM sodium phosphate buffer pH 7.2
250 mM NaCl	
50 mM KCl	100 mM KCl
2 mM MgCl2	2 mM MgCl2
1 mM DTT	1 mM DTT
	1.5 mM desthiobiotin

tRNA Extraction for *in vitro* Biochemistry

A pre-culture of the desired strain was incubated overnight at 37°C with vigorous shaking. 5 mL of pre-culture was diluted in 100 ml of pre-warmed rich medium (2xYT) and the cells ware incubated at 37°C (1-2h, 220 rpm). When OD_{600}=0.5~1.0, cells were pelleted (5 min, 5,000 rcf), the supernatant was discarded and the cell pellet was resuspended in 1.5 mL of Buffer D (composition below) and transferred to a 2 ml tube.

The cells were pelleted (2 min, 5,000 rcf), the supernatant completely removed and the cell pellet resuspended in 900 µL of Buffer D, then 100 µL of liquefied phenol (90% phenol in water from Sigma Aldrich, cat. No. P9346) was added to the cells and the lysis was performed on a head-over-tail rotation (15 min, 15 rpm, room temperature).

The lysed cells were centrifuged (25 min, >20,000 rcf, 4°C). ~1 mL of supernatant containing the tRNAs was recovered and transferred to a new tube. The solution should not display phase separation. 1 mL of chloroform was added to the solution and the tube was thoroughly vortexed for 1 min until a cloudy emulsion was formed.

The emulsion was separated by centrifuging (1 min, >20,000 rcf) and the top aqueous layer contains the tRNAs and was recovered (~950 µl).

The extracted tRNA solution was deacylated by addition of 56 µL of NaOH 300 mM (42°C, 1h), then neutralized with 56 µL of NaOAc 3 M pH 5. The solution was brought to 70% EtOH using absolute ethanol and incubated at 4°C for 1h, then the precipitated tRNAs were centrifuged (>16,000 rcf, 20 min). The supernatant was removed and the pellet was left to dry, then resuspended in water and abundantly washed with water by means of Sartorius Vivaspin® 2 with 3kDa cutoff spin column until the solution was totally clear.

tRNAs are diluted to a final concentration of 2 µg/µL and frozen at -20°C.

Buffer D

NaOAc 50 mM pH 5	NaCl 150 mM	$MgCl_2$ 10 mM	EDTA 0.1 mM

In vitro Aminoacylation

^{14}C-labelled amino acids were purchased from Moravek Inc. (500 µl solution, 100 µCi/mL, variable specific concentration). Each amino acid was then diluted with cold amino acid to a final concentration of 75 µCi/mL and a specific activity of 60 mCi/mmol (1.25 mM, Amino Acid Stock 5x). In the case of histidine, the amino acid stock 5x was prepared by diluting the ^{3}H-labelled amino acid, purchased by the same supplier at a concentration of 1.0 mCi/mL, to a final concentration of 0.75 mCi/mL and a specific concentration of 600 mCi/mmol (1.25 mM). The reaction was performed in Aminoacylation buffer 1x (composition below). tRNA Stock was prepared at 2 µg/µl, while the Enzyme Stock was prepared at concentrations between 4 and 12.5 µM. Before the reaction is started, Whatman glass microfiber filters GF/B (21 mm, CAT No. 1821-021) were wetted with 250 µL of cold trichloroacetic acid 5%. For each enzyme, a 2x Master Mix (2xMM) was assembled as follows:

Aminoacylation Buffer 10x	**DTT 20 mM**	**Amino Acid Stock 5x**	**Enzyme Stock**	**ddH_2O**
32 µl	12.8 µL	64 µL	3.84 µL	15.36 µL

30 µL of the 2xMM was mixed with 37.5 µL of tRNA Stock 2x. 9 µL was spotted on a wet glass filter as time 0 min. The reaction was started by adding 6.5 µL of ATP 10 mM, then 10 µL aliquots were taken at time 0.5, 1, 2, 3 and 5 min and spotted on filters. The filters were transferred to a vacuum filtration apparatus and washed with 1 ml of 5% TCA; then 4 ml of 1% TCA, then 1 ml of 70% ethanol to remove the unbound amino acid. The filters were then transferred to the appropriate vial for liquid scintillation counting. To measure total counts, 10 µL of each reaction was transferred to a dry filter and directly transferred to a vial for liquid scintillation counting without any washing.

Aminoacylation buffer (10x)

200 mM HEPES pH 7.2	500 mM KCl	100 mM $MgCl_2$

tRNA Purification and Amino Acid Analysis

A biotinylated DNA probe (100 µM) was added to the tRNA solution to a final concentration of 0.1-1 µM. The mixture was heated to 75°C for 10 min to anneal the probe to the target tRNA. The extract was then incubated on ice and the DNA/tRNA hybrid captured on high-capacity streptavidin-coated agarose (a volume of beads equal to twice the volume of 100 µM biotinylated probe used for annealing was added). The solution was shaken (30 min, 4˚C), then the beads were recovered and transferred to a 1 mL empty chromatographic spin column (Bio-rad, cat. No. 7326207). The beads were washed 5 times with 100 mM ammonium acetate pH 5, then 3 times with 20 mM ammonium acetate pH 5.

To hydrolyse the amino acid from tRNA, 200 µL of 20 mM ammonium carbonate pH 9.6 was added and the solution incubated (20 min, 37˚C) before collection. To maximize recovery of the amino acid, the beads were further washed with 200 µL of water, 200 µL of 20 mM ammonium bicarbonate pH 3 twice, then 200 µL of water twice. All the washes were combined and lyophilised until dry. The amino acid was dissolved in 40 µL of water/methanol (40:60), the solution was centrifuged (30 min, 21,000 rcf, 4˚C) to sediment precipitation. The top 20 µL of the solution was transferred into a 250 µL glass insert (Agilent) for MS analysis.

All amino acid samples were analysed on an Agilent 1260 Infinity equipped with an Agilent 6130 Quadrupole LC-MS unit. A HILIC-Z column (4.6x150 mm) equipped with a guard column (Agilent) with 0.5 mL/min flow rate was used to elute amino acids. Buffer A and Buffer B (compositions below) were used for RP-HPLC. 5 µL of amino acid-containing solution was injected and eluted using a linear gradient of 100%:0% Buffer B:buffer A to 70%:30% Buffer B:Buffer A over 10 min at 30˚C. The mass spectrometer was set to selected-ion monitoring (SIM) mode. Pure amino acids, purchased from Sigma Aldrich, were run as standards for comparison.

Buffer A

10 mM NH_4HCO_2 in H_2O

Buffer B

CH_3CN: H_2O 9:1 (v/v), 10 mM NH_4HCO_2

GFP Expression and Mass Spectrometry

The gene encoding sfGFP is contained in a plasmid with the p15A origin of replication under the L-arabinose-inducible P_{BAD} promoter. The resulting protein has the following sequence, with the * at position 150 corresponding to the amber stop codon used for ncAA incorporation:

```
  1 MVSKGEELFT GVVPILVELD GDVNGHKFSV RGEGEGDATN GKLTLKFICT TGKLPVPWPT
 61 LVTTLTYGVQ CFSRYPDHMK RHDFFKSAMP EGYVQERTIS FKDDGTYKTR AEVKFEGDTL
121 VNRIELKGID FKEDGNILGH KLEYNFNSH* VYITADKQKN GIKANFKIRH NVEDGSVQLA
181 DHYQQNTPIG DGPVLLPDNH YLSTQSVLSK DPNEKRDHMV LLEFVTAAGI THGMDELYKG
241 SHHHHHH-
```

E. coli DH10b cells harbouring this sfGFP-containing plasmid were transformed using heat shock (42°C, 50 s, 150 ng DNA, 25 µl of cells) with a plasmid that constitutively expresses tRNA/synthetase pair (pRST). The plasmid contained the pMB1 origin of replication. Following transformation, cells were recovered for 1h (1 mL SOC medium, 37°C, 850 rpm). After recovery, the transformation was diluted 100 fold in 2xYT containing antibiotics for selection of transformants, 0.2% L-arabinose and the appropriate ncAA (2 mM) when required. After overnight expression, 5 to 50 mL of cells were harvested by centrifugation (5,000 rcf, 10 min), the supernatant was discarded and the pellet was resuspended in 1 mL of lysis buffer (composition below) and incubated for 15 min on head-over-tail rotation at room temperature. The crude lysate was centrifuged at >16,100 rcf for 20 min at 4°C, the supernatant was recovered and 50-100 µl of 1:1 slurry of Ni-NTA Agarose beads (Qiagen) was added. The beads were recovered by centrifugation at 200 rcf, washed 5 times with 1 mL of wash buffer (composition below), then sfGFP was eluted using 50 µL of elution buffer (composition below).

Mass spectra of all protein samples were acquired on an Agilent 1200 LC-MS system equipped with a 6130 Quadrupole spectrometer. A Phenomenex Jupiter C4 column (150×2 mm, 5 µm) was used to elute proteins. Buffer A and Buffer B (compositions below) was used for RP-HPLC. Mass spectra were acquired in the positive mode and analysed by the MS Chemstation software (Agilent Technologies). The deconvolution program provided in the software was used to obtain the entire mass spectra.

Lysis buffer

BugBuster™ Protein Extraction Reagent	20 mM Tris pH 8.0	500 mM NaCl	40 mM imidazole

Wash buffer

20 mM Tris pH 8.0	500 mM NaCl	40 mM imidazole

Elution buffer

50 mM Tris pH 8.0 | 50 mM NaCl | 300 mM imidazole

Buffer A

0.2% HCOOH in H_2O

Buffer B

0.2% HCOOH in CH_4CN

GFP Total Mass

The table below summarises the total mass of the reporter sfGFP^{150*} depending on the amino acid present at position 150.

Amino acid in position 150		Mass after flurophore maturation	Mass after flurophore maturation (-Met)
gly	G	27770,27	27639,07
ala	A	27784.30	27653.10
ser	S	27800.30	27669.10
pro	P	27810,34	27679,14
val	V	27812,35	27681,15
thr	T	27814,32	27683,12
cys	C	27816,36	27685,16
ile	I	27826,38	27695,18
leu	L	27826,38	27695,18
asn	N	27827,32	27696,12
asp	D	27828,31	27697,11
gln	Q	27841,35	27710,15
lys	K	27841,39	27710,19
glu	E	27842,33	27711,13
met	M	27844,41	27713,21
his	H	27850,36	27719,16
phe	F	27860.40	27729.20
arg	R	27869,41	27738,21
tyr	Y	27876.40	27745.20
O-methyl-L-tyrosine		27890,42	27759,22
trp	W	27899,43	27768,23
p-N_3-L-phenylalanine		27901.40	27770.20
N^{ε}-boc-L-lysine		27941,31	27810,11
p-iodo-L-phenylalanine		27986,29	27855,09

GFP Expression for Fluorescence Quantification

E. coli DH10b cells containing the sfGFP gene with an in-frame amber stop codon at position 150 (see above) were transformed with a pWK or pRST plasmid using heat shock. Recovery medium was added (1 ml, SOC medium) and the transformation was divided into 3 aliquots (biological replicates). After recovery (1h, 37°C, 850 rpm), 50 µl of each transformation was added to a different well of a 24 deep well plate (Riplate®SW 24, PP, 10 ml) containing 5 ml of 2xYT medium supplemented with the appropriate antibiotics for selection, 0.2% arabinose to induce expression of GFP and 2 mM of ncAA, as necessary. Cells were incubated at 37°C (22h, 220 rpm), then 500 µL of medium was transferred into clear-bottom 24 well plates and the GFP fluorescence in this standard volume was measured. Measurements were performed on each sample grown in individual wells. In each graph, individual data points are shown together with the standard deviation.

Site-Saturation Mutagenesis

Site saturation mutagenesis on aaRSs or tRNAs was carried out by enzymatic inverse PCR (eiPCR)[124]. Primers were designed to contain the BsaI or SapI restriction sites. PCRs were performed using Q5® Hot Start High-Fidelity DNA Polymerase (NEB) using the protocol below (indicated for 20 µL reactions).

DNA template (1 ng/µl)	Primers Forw+ Rev (5 µM each)	dNTPs (10 mM each)	5x Q5 Buffer	Q5 Polymerase	H_20
1 µL	0.8 µL	0.4 µL	4 µL	0.2 µL	13.6 µL

Followed by the thermal cycles as follows:

		28x cycles						
95°C	→	**95°C**	→	**56/58°C**	→	**72°C**	→	**16°C**
30 s		10 s		10 s		30 s/kpb		∞

Alternatively, PrimeSTAR® GXL DNA Polymerase (TaKaRa) was used using the protocol indicated below:

DNA template (1 ng/µl)	Primers Forw+ Rev (5 µM each)	dNTPs (10 mM each)	5x Q5 Buffer	Q5 Polymerase	H_20
1 µL	0.8 µL	0.4 µL	4 µL	0.4 µL	13.4 µL

Followed by the thermal cycles as follows:

		28x cycles						
95°C	→	**95°C**	→	**58/68°C**	→	**68°C**	→	**16°C**
30 s		10 s		10 s		60 s/kpb		∞

Following amplification, the PCRs were purified by adding 3 volumes of buffer PB (Qiagen) before using the QIAprep 2.0 Spin Miniprep Columns (Qiagen) to bind the DNA. The column were washed with PE buffer (Qiagen), then the DNA was eluted using EB buffer (Qiagen) and its concentration measured using a nanodrop.

Restriction digest was performed as follows (indicated for a 50 µL reaction):

DNA	**BsaI/SapI**	**DpnI**	**CutSmart**	**H_20**
50 ng/µL	1 µL	0.5 µL	5 µL	to 50 µL

After 2h incubation at 37°C, the DNA was purified as indicated above, then ligation was performed as indicated below (indicated for a 20 µL reaction):

DNA	**T4 ligase**	**T4 ligase buffer**	**H_20**
10 ng/µL	1 µL	2 µL	to 20 µL

The reaction volumes indicated above were scaled based on the total amount of DNA processed. Following overnight incubation at 18°C, the DNA was purified as indicated above and transformed into electrocompetend *E. coli* DH10b cells. Libraries were produced to contain >2x10^8 individual transformants.

Mutagenesis by Error-Prone PCR

The gene coding for the synthetase targeted for evolution was amplified by PCR using primer immediately flanking the CDS. The DNA obtained was subsequently purified and used as a template for random mutagenesis by error-prone PCR, which was carried out using the GeneMorph II Random Mutagenesis Kit (Agilent Technologies). Amplification was carried out according to the manufacturer's instructions to achieve medium mutation frequency (4.5-9 mutations/kb). 30 cycles were performed with 250 ng of initial target DNA. Random mutagenesis libraries were generated by assembly of the mutagenised synthetase with the plasmid backbone, which was amplified using Q5® Hot Start High-Fidelity DNA Polymerase (NEB, Cat. No. M0493S). Assembly was performed for 1h at 50°C using NEBuilder HiFi DNA Assembly Master Mix, followed by transformation to homemade

electrocompetent cells. The transformation efficiency was >$3x10^7$.

aaRS Library Selections

pRST plasmids containing the synthetase of interest together with its cognate tRNA were used to build the libraries. Ten aliquots of 1 µg of library DNA was transformed into ten aliquots of 100 µL of homemade electrocompetent *E. coli* DH10b cells containing an appropriate reporter plasmid (e.g.: p15A-*cat*^{112}*-*gfp*^{150}* , p15A-*cat*^{112}*-*gfp*^{67}* or p15A-*cat*^{112*}-*gfp*^{67*}*e2crimson*). To ensure high electroporation efficiency, electro-competent cells were freshly prepared by diluting 25 fold an overnight pre-culture of the appropriate cells into 2xYT medium containing the appropriate antibiotic. The cells were grown to OD_{600}=0.6~0.8. Cells were then harvested and resuspended in ice-cold 10% glycerol in water; this procedure is repeated 3 times. Finally, the cells were harvested and the cell pellet resuspended with 10% glycerol in water (2 times the pellet volume). After electroporation, cells were recovered (1h at 37°C, SOC medium) and then were incubated overnight (37°C, 1 L of 2xYT, 220 rpm) in medium contained the appropriate antibiotics (commonly, tetracycline to select for the reporter plasmid and spectinomycin to select for the library plasmid). An appropriate dilution of this medium was plated on agar plates to ensure the transformation had yielded >$5x10^8$ individual transformants (e.g.: >500 colonies on the 10^{-6} dilution plate). The overnight culture of library-containing transformants was then diluted 25-fold in 2xYT and allowed to grow to OD_{600}=1; if necessary, the appropriate ncAA was the added to a concentration of 2 mM. When OD_{600}=1 was reached, 2 mL of cells were plated on 25x25 cm^2 agar plates containing the appropriate antibiotics to select for the markers on the plasmids, together with chloramphenicol, 0.2% arabinose to induce expression of the fluorescent markers (e.g. sfGFP or E2Crimson), and the appropriate ncAA (2 mM) if needed. The chloramphenicol concentration was chosen in accordance to the desired level of activity of the synthetase (range: 25-500 µg/mL). The plates were incubated at 37°C overnight. For selections using the reporter plasmids p15A-*cat*^{112}*-*gfp*^{150}*or p15A-*cat*^{112}*-*gfp*^{67}*, surviving colonies were illuminated with blue light (488 nm) to test for GFP production. For selections using the reporter plasmid p15A-*cat*^{112*}-*gfp*^{67*}*e2crimson*, surviving colonies were illuminated with blue light (488 nm) to test sfGFP production and with red light (635 nm) to test for E2Crimson production. Individual colonies displaying the correct fluorescence pattern were grown and characterised, or the surviving cells were pooled together, the DNA extracted and transformed again for another round of selection or for a round of activity screening.

tRNA Library Selections for Improved Orthogonality and Activity

tRNA selections were performed using a two-plasmid system. A pKW plasmid, containing the tRNA was used to build the library via site-saturation mutagenesis. A second plasmid was a modified version of the appropriate reporter plasmid (e.g.: p15A-*cat*^{112*}-*gfp*^{150*} or p15A-*cat*^{112*}-*gfp*^{67*}) which additionally contained the synthetase.

1 µg of library DNA was transformed into homemade electro-competent *E. coli* DH10b cells (0.1 mL) containing the appropriate synthetase-containing reporter plasmid. Following recovery (1h at 37°C, SOC medium 220 rpm, the cells were incubated overnight (37°C, 100 mL of 2xYT, 220 rpm). The medium contained the appropriate antibiotics (commonly, tetracycline to select for the reporter and spectinomycin to select for the library plasmid). An appropriate dilution (e.g.: 10^{-5}) of the transformation was plated on agar plated to ensure the transformation efficiency was enough to yield >5x10^7 individual transformants. The overnight culture was then diluted 25-fold in 2xYT and allowed to grow to OD_{600}=1. When OD_{600}=1 was reached, 2 mL of cells were plated on 25x25 cm^2 agar plates containing the appropriate antibiotics to select for the markers on the plasmids, together with chloramphenicol and 0.2% arabinose to induce expression of GFP. The chloramphenicol concentration was chosen in accordance to the desired level of activity of the synthetase (range: 25-500 µg/mL). The plates were incubated at 37°C overnight. Surviving colonies were illuminated with blue light to test for GFP production. DNA was extracted from either individual GFP-positive colonies, or from GFP-positive colonies pooled together.

The reporter plasmid was selectively digested with an appropriate restriction enzyme and T5 exonuclease (NEB, cat. No. M0363S) treatment. The now isolated tRNA library plasmid was purified again and transformed into cells containing the appropriate selection plasmid (e.g.: p15A-*cat*^{112*}-*gfp*^{150*} or p15A-*cat*^{112*}-*gfp*^{67*}) lacking the cognate synthetase, and the transformants are plated on agar plates containing 0.2% arabinose for GFP screening in the absence of chloramphenicol. GFP-negative colonies contain a variant of the tRNA which is selectively aminoacylated to a desired level in the presence of the synthetase and not in its absence, thus being more active and orthogonal.

Chapter V – Appendix

tRNA Experimentally Tested

Below, the compete list of tRNA sequences experimentally analysed is displayed. Each tRNA is indicated by an arbitrary identifier, then the species of origin is indicated, together with the anticodon sequence and the unique ID assigned to that tRNA in the tRNA-DB-CE. The tRNA sequence indicated includes the 3'-CCA end (positions 74→76). The tREX probes designed to detect each tRNA, whose sequence is indicated, were purchased with the cyanine 5 fluorophore bound at their 3'-end.

Arg_01from *Rothia mucilaginosa*, anticodon: CCG, Seq. ID: C10110712

Seq.: GCCTCTGTAGCTCAGCGGATTAGAGCACCAGTTTCCGGTACTGGGGGTCGAAGGTTCGAATCCTTTCAGGGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCCCTGAAAGGATTCGAACCTTCGACC

Arg_02from *Corynebacterium accolens*, anticodon: CCG, Seq. ID: W09122376

Seq.: GTCTCCGTAGCTCAGCGGATTAGAGCATCGGTTTCCGGTACCGAAGGTCGCAGGTTCGATCCCTGTCGGGGACACCA

tREX probe:aaaaaaaaaaaaaTGGTGTCCCCGACAGGGATCGAACCTGCGACC

Arg_03from *Corynebacterium aurimucosum*, anticodon: CCG, Seq. ID: W09127825

Seq.: GCCCTTGTAGCTCAGCGGATTAGAGCATCGGTTTCCGGTACCGAAGGTCGCAGGTTCGATCCCTGTCAGGGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCCCTGACAGGGATCGAACCTGCGACC

Arg_04from *Capnocytophaga sp. oral taxon 329*, anticodon: CCG, Seq. ID: W11301568

Seq.: GGCTCCGTAGCTTAGCTGTATAGAGCGTCAGATTCCGGTTCTGAAGGTCGAGGGTTAGAATCCCTCCGGGGTCACCA

tREX probe:aaaaaaaaaaaaaTGGTGACCCCGGAGGGATTCTAACCCTCGACC

Arg_05from *Fluviicola taffensis*, anticodon: TCG, Seq. ID: C11300221

Seq.: GGCTCCGTAGCTCAGCTGTATAGAGCACTGGATTTCGGCTCCAGCGGTCGGGAGTTAGAATCTCTCCGGGGTCACCA

tREX probe:aaaaaaaaaaaaaTGGTGACCCCGGAGAGATTCTAACTCCCGACC

Arg_06from *Amycolatopsis azurea*, anticodon: TCG, Seq. ID: W131064644

Seq.: CATGGAGTGGATCAGCGGCAGATCGCCCGGCTTCGGTCCGGGAGGGCGCGGGTTCGAGTCCCGCCTCCATGTCCA

tREX probe:aaaaaaaaaaaaaTGGACATGGAGGCGGGACTCGAACCCGCGCCC

Arg_07from *Caldilinea aerophila*, anticodon: GCG, Seq. ID: C121001241

Seq.: GTCCCCGTAGCTCAGTGGATGAGAGCGCTTGGTTGCGGTCCAAGAGGTCAGAGGTTCGAGTCCTCTCGGGGATGCCA

tREX probe:aaaaaaaaaaaaaTGGCATCCCCGAGAGGACTCGAACCTCTGACC

Arg_08from *Enterococcus faecalis*, anticodon: TCG, Seq. ID: W131189827

Seq.: TGGCGTGTAGCATTGTGGTAATGCAACTTATTTCGTGTGAGATAAGATGCGGGTTCGAATCCTGTCACGCCAACCA

tREX probe:aaaaaaaaaaaaaTGGTTGGCGTGACAGGATTCGAACCCGCATCTT

Arg_09from *Weeksella virosa*, anticodon: TCG, Seq. ID: C11300548

Seq.: GGCCTCGTAGCTCAACTGTATAGAGCACTGGATTTCGGCTCCAGCGGTTGGGGGTTAGAATCCCTCCGAGGTCACCA

tREX probe:aaaaaaaaaaaaaTGGTGACCTCGGAGGGATTCTAACCCCCAACC

Arg_10from *Corynebacterium striatum*, anticodon: CCG, Seq. ID: W09122422

Seq.: GCCTCCGTAGCTCAGCGGATTAGAGCATCGGTTTCCGGTACCGAAGGTCGCAGGTTCGATCCCTGTCGGGGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCCCCGACAGGGATCGAACCTGCGACC

Asn_01from *Rivularia sp. PCC 7116*, anticodon: ATT, Seq. ID: C131004222

Seq.: TACATACTAGCTCATTGGTAGAGCATTCGGCTATTAACCGACTGGTAGTAGGTTCAAATCCTGCGTATGAACCCA

tREX probe:aaaaaaaaaaaaaTGGGTTCATACGCAGGATTTGAACCTACTACC

Asn_02from *Chamaesiphon minutus*, anticodon: ATT, Seq. ID: C131004558

Seq.: CAAAAGTTAGCTCATTAGGTAGAGCAGTTGACTATTAATCAACGTGTAACAGGTTCGATTCCTGTACTTTTCACCA

tREX probe:aaaaaaaaaaaaaTGGTGAAAAGTACAGGAATCGAACCTGTTACA

Asn_03from *Oscillatoria nigro-viridis*, anticodon: ATT, Seq. ID: C131004769

Seq.: ACAAAGTTAGCTCAATTGGTAGAGCGATCGACTATTAATCGATTGGTAACAGGTTCAATTCCTGTACTTTAAACCA

tREX probe:aaaaaaaaaaaaaTGGTTTAAAGTACAGGAATTGAACCTGTTACC

Asn_04from *Fischerella sp. JSC-11*, anticodon: ATT, Seq. ID: W11181271

Seq.: TACATACTAGCTCAATTGGTAGAGCAGTCGGCTATTAACCGATGGGTAACAGGTTCGATTCCTGTGTATAAACCCA

tREX probe:aaaaaaaaaaaaaTGGGTTTATACACAGGAATCGAACCTGTTACC

Asn_05from *Bacillus cereus*, anticodon: GTT, Seq. ID: W121028747

Seq.: TCCGCAGTAGCTCAGTGGTAGAGCTAAGAGCTATCGGCTGTTAACCGATCGGTCGTAGGTTCGAGTCCTACCTGCGGAGCCA

tREX probe:aaaaaaaaaaaaaaaaTGGCTCCGCAGGTAGGACTCGAACCTACGACCGATCGGT

Asn_06from *Lactobacillus hominis*, anticodon: ATT, Seq. ID: W141244219

Seq.: AAGCCTATAGTTCAATTGGTAGAACGCTCTACTATTACTGAAGTAGATATGAGGGTTCGAATCCCTTTAGGCTTACCA

tREX probe:aaaaaaaaaaaaaaTGGTAAGCCTAAAGGGATTCGAACCCTCATATCT

Asn_07from *Serratia marcescens*, anticodon: ATT, Seq. ID: C151100364

Seq.: ACGTCATTAGCTCAGCAGGAAGAGCGGCAGCGAAATATTCGTTGGCGGACTGGGGTTCGATTCCTCGATGACGTTCCA

tREX probe:aaaaaaaaaaaaTGGAACGTCATCGAGGAATCGAACCCCAGTC

Asn_08from *Salmonella enterica*, anticodon: GTT, Seq. ID: W1510481293

Seq.: TCCTCTGTAGTTCAGTCGGTAGAACGGGCGGACTGTTAATCCGTATGTCACTGGTTCGAGTCCAGTCAGAGGAGCCA

tREX probe:aaaaaaaaaaaaaaTGGCTCCTCTGACTGGACTCGAACCAGTGACAT

Asn_09from *Klebsiella pneumoniae*, anticodon: ATT, Seq. ID: W1511345087

Seq.: GGGTCGTTAGCTCAGTTGGTAGAGCAGTTGACTATTAATCAATTGGTCGCAGGTTCGAATCCTGCACGACCCACCA

tREX probe:aaaaaaaaaaaaaaTGGTGGGTCGTGCAGGATTCGAACCTGCGACC

Asn_10from *Providencia rettgeri*, anticodon: ATT, Seq. ID: W141022179

Seq.: CCGCCATTAACTCAACTGGAAGAGTATTTAGCCTTATTAAGGCTAAGAGTCGAGGTTCGATGCCTCGATGGCGGACCA

tREX probe:aaaaaaaaaaaaaaTGGTCCGCCATCGAGGCATCGAACCTCGACTC

Asp_01from *Allokutzneria albata*, anticodon: GTC, Seq. ID: W141766821

Seq.: CCTGATGTAGATCAGCGGTAGATCGCCCGGCTGTCAACCGGGAGGGCGCGAGTTCGAATCTCGCCATCAGGTCCA

tREX probe:aaaaaaaaaaaaaaTGGACCTGATGGCGAGATTCGAACTCGCGCCC

Asp_02from *Bradyrhizobium sp. OHSU_III*, anticodon: GTC, Seq. ID: W141063677

Seq.: GCCTCCGGACCCAGCTGGCGACGGGACTCGACTGTCGATCGAGCATTGGCGGGTTCGATTCCCGCCGGAGGCGCCA

tREX probe:aaaaaaaaaaaaaaTGGCGCCTCCGGCGGGAATCGAACCCGCCAAT

Asp_03from *Chryseobacterium angstadtii*, anticodon: ATC, Seq. ID: W1511685144

Seq.: GAAACGTTAGCTCAATGGAAGAGCTTTGACTTATCTGAACAGAGGCTGCGGGTTCGAATCCCGCACGTTTCTCCA

tREX probe:aaaaaaaaaaaaaaTGGAGAAACGTGCGGGATTCGAACCCGCAGCC

Asp_04from *Thiobacillus denitrificans*, anticodon: GTC, Seq. ID: W131160302

Seq.: GCCTCGTTAACTCAGCGGCAGAGTGCCCGCCTGTCTAGCGGGAAGCCAGGGGTTCAAGTCCCCTACGAGGCGCCA

tREX probe:aaaaaaaaaaaaaaTGGCGCCTCGTAGGGGACTTGAACCCCTGGCT

Asp_05from *Mycobacterium phage PegLeg*, anticodon: GTC, Seq. ID: PHG14102396

Seq.: CCCGATGTCATCTAGTGGACCAGGATGCCGCCCTGTCGAGGCGGTCACGCGAGTTCGAATCTCGTCGTCGGGACCA

tREX probe:aaaaaaaaaaaaTGGTCCCGACGACGAGATTCGAACTCGCGTG

Asp_06from *Listeria phage vB_LmoM_AG20*, anticodon: GTC, Seq. ID: PHG14101575

Seq.: GTGCGTATGATATAATGGCTATTATACTCGACTGTCTATCGAGAAATAGGGGTTCAATTCCCCTTACGTGCGCCA

tREX probe:aaaaaaaaaaaaTGGCGCACGTAAGGGGAATTGAACCCCTATT

Asp_07from *Variovorax paradoxus*, anticodon: GTC, Seq. ID: C09111138

Seq.: GATCCGTTAACTCAGCGGCCAGAGTGCCTCCCTGTCCAGGAGGAAGCCAAGGGTTCAACTCCCTTACGGATCGCCA

tREX probe:aaaaaaaaaaaaaaTGGCGATCCGTAAGGGAGTTGAACCCTTGGCT

Asp_08from *Bradyrhizobium oligotrophicum*, anticodon: GTC, Seq. ID: C132000002

Seq.: GCCCATGTCCCCATCTGGCGAAGGGACCGGACTGTCGATCCGGAAAGGCGGGTTCGATTCCCGTCATGGGCGCCA

tREX probe:aaaaaaaaaaaaTGGCGCCCATGACGGGAATCGAACCCGCCTT

Asp_09from *Sorangium cellulosum*, anticodon: GTC, Seq. ID: C08012733

Seq.: GCCCTCGTATGCATCTGGTGAGGCAGCCTGATTGTCGATCAGGTGAGGGGAGTTCGATCCTCCTCGGGGGCGCCA

tREX probe:aaaaaaaaaaaaTGGCGCCCCCGAGGAGGATCGAACTCCCCTC

Cys_01from *Enterococcus caccae*, anticodon: GCA, Seq. ID: W131028011

Seq.: GATGGTATAGCCAAGTGGTAGGGCACAGGTCTGCAAAACCTTGAGCATCGGTTCAAACCCGATTACCATCTCCA

tREX probe:aaaaaaaaaaaaTGGAGATGGTAATCGGGTTTGAACCGATGC

Cys_02from *Niastella koreensis*, anticodon: GCA, Seq. ID: C123000104

Seq.: GAGCAGGTACTCAATACTGACATGGAGATTGGTCTGCAAAACCAATTCAATATGGGTTTAATTCCCATCCTGCTCTCCA

tREX probe:aaaaaaaaaaaaaTGGAGAGCAGGATGGGAATTAAACCCATATTGA

Cys_03from *Acinetobacter lwoffii*, anticodon: GCA, Seq. ID: W121047593

Seq.: GCATGTATGGGTGAGGTGGATTAGCCAGCGGACTGCAACTCCGTTTATACAGGTTCGAGTCCTGTTACATGCTCCA

tREX probe:aaaaaaaaaaaaTGGAGCATGTAACAGGACTCGAACCTGTATA

Cys_04from *Chitinophaga pinensis*, anticodon: GCA, Seq. ID: C10103731

Seq.: TGAGATGTGGTCGAATCTGGAATGACGCCCCCTTGCAACGGGGGAAAAATAAGTTCGAATCTTATCTTCTCTCCCA

tREX probe:aaaaaaaaaaaaTGGGAGAGAAGATAAGATTCGAACTTATTTT

Cys_05from *Moorea producens*, anticodon: GCA, Seq. ID: W11143160

Seq.: GTCCAGGTCGCCTAATGGTAGGGCATTGGTCTGCAAAACCGATTGTGTGAGTTCAATTCTCACCCTGGACTCCA

tREX probe:aaaaaaaaaaaaTGGAGTCCAGGGTGAGAATTGAACTCACACA

Cys_06from *Roseburia inulinivorans*, anticodon: GCA, Seq. ID: W09300492

Seq.: GATGCTGTAGCCGCTGGTAGGGCATAGGTTTGCAAAACCGAAGTTGGTGGTTCGAATCCGCCCAGTATCACCA

tREX probe:aaaaaaaaaaaaTGGTGATACTGGGCGGATTCGAACCACCAAC

Cys_07from *Calothrix sp. PCC 7507*, anticodon: GCA, Seq. ID: C131010320

Seq.: GTCCAGGTCGCCTAATGGTAGGGTATTGGTCTGCAAAACCGACTGTGCGAGTTCAATTCTCGCCCTGGACTCCA

tREX probe:aaaaaaaaaaaaTGGAGTCCAGGGCGAGAATTGAACTCGCACA

Cys_08from *Scytonema hofmannii*, anticodon: GCA, Seq. ID: W131065815

Seq.: GTCCAGGTCGCCTAACGGTAGGGCGTTGGTCTGCAAAACCGACTGTGCGAGTTCAATTCTCGCCCTGGACTCCA

tREX probe:aaaaaaaaaaaaTGGAGTCCAGGGCGAGAATTGAACTCGCACA

Gln_01from *Candidatus Solibacter usitatus*, anticodon: TTG, Seq. ID: C024933

Seq.: TCCCCGGTCGTCTAATGGTAGGACAGCGGCCTTTGGAGCCGTGAATCGTGGTTCGAATCCATGCCGGGGAGCCA

tREX probe:aaaaaaaaaaaaTGGCTCCCCGGCATGGATTCGAACCACGATT

Gln_02from *Mycobacterium abscessus*, anticodon: CTG, Seq. ID: W1510544614

Seq.: CAAACACTAGCTCAATTGGCAGAGCAATCGTTTCTGGAACGGTAGGTCGCCGGTTCAAGTCCGGCGTGTTTATCCA

tREX probe:aaaaaaaaaaaaaTGGATAAACACGCCGGACTTGAACCGGCGACC

Gln_03from *Propionimicrobium lymphophilum*, anticodon: TTG, Seq. ID: W131002891

Seq.: TCCCCGGTTGGTCTGCTGGTAGGGCCCGCCTGACTTTGAATCAGGATTTAGCGACGCAGGTTCGATTCCTGCCCGGGGAACCA

tREX probe:aaaaaaaaaaaaaaaTGGTTCCCCGGGCAGGAATCGAACCTGCGTCGCTAAA

Gln_04from *Ilumatobacter nonamiensis*, anticodon: TTG, Seq. ID: W131240735

Seq.: TCCCGGGTCGTCTAACGGTAAGACAGCGGTTTTTGGTGCCGAGAATAGGGGTTCGATTCCTCTCCCGGGAACCA

tREX probe:aaaaaaaaaaaaTGGTTCCCGGGAGAGGAATCGAACCCCTATT

Gln_05from *Bradyrhizobium sp. Tv2a-2*, anticodon: TTG, Seq. ID: W141150373

Seq.: ACCCGCTTGGTCCAGTGGTCTAAGACGTCGGACTTTGACTCCGAAGACAGTCGTTCAAATCGACTAGCGGGTGCCA

tREX probe:aaaaaaaaaaaaTGGCACCCGCTAGTCGATTTGAACGACTGTC

Gln_06from *Flavihumibacter petaseus*, anticodon: CTG, Seq. ID: W1510019463

Seq.: AGTTGCGTAGTGCAAAGGCAGCACATCTGAATCTGGCTCAGAAGATGTTGGTTCGAATCCAGCCGCAACGACCA

tREX probe:aaaaaaaaaaaaTGGTCGTTGCGGCTGGATTCGAACCAACATC

Gln_07from *Chryseobacterium tenax*, anticodon: TTG, Seq. ID: W131220961

Seq.: AGACCTATGGTGTAGCGGTAACACTACTGTTTTTGGTGCAGTCATCTGGGGTTCGAATCCCTGTGGGTCTACCA

tREX probe:aaaaaaaaaaaaTGGTAGACCCACAGGGATTCGAACCCCAGAT

Gln_08from *Truepera radiovictrix*, anticodon: CTG, Seq. ID: C10113751

Seq.: TCAGGAATGGTGTAACGGTAGCACGACGCACTCTGGATGCGTTAGTCCTGGTTCGAATCCAGGTTCCTGAGCCA

tREX probe:aaaaaaaaaaaaTGGCTCAGGAACCTGGATTCGAACCAGGACT

Gln_09from *Rhodopirellula baltica*, anticodon: CTG, Seq. ID: C016619

Seq.: GCCGGAGTAGCAGTTGTTGGTCGCTGCACCTGACTCTGAATCAGGAGTTCATTGGTTCGATTCCAATCTCCGGTGCCA

tREX probe:aaaaaaaaaaaaaTGGCACCGGAGATTGGAATCGAACCAATGAAC

Gln_10from *Verrucomicrobia bacterium SCGC AAA164-O14*, anticodon: CTG, Seq. ID: W141240493

Seq.: AGCCCGGTAGTGTAGTGGCAAGCATGGGGGATTCTGGATCCCTTGACCTCGGTTCGAGTCCGAGCCGGGCTACCA

tREX probe:aaaaaaaaaaaaTGGTAGCCCGGCTCGGACTCGAACCGAGGTC

Glu_01from *Lactobacillus delbrueckii*, anticodon: CTC, Seq. ID: C011433

Seq.: CACCCGTTGGTCAAGTGGTTAAGACGCTACCCTCTCAAGGTGGAGTCATGAGTTCAATTCTCGTACGGGTGACCA

tREX probe:aaaaaaaaaaaaTGGTCACCCGTACGAGAATTGAACTCATGAC

Glu_02from *Oscillatoria nigro-viridis*, anticodon: TTC, Seq. ID: C131004735

Seq.: AGTGCTGTAGGCAAATGGTTTAAGCCGCTTGGTTTTCACCCAAGTGATTGCGGGTTCAACTCCCGTCAGCACTTCCA

tREX probe:aaaaaaaaaaaaaTGGAAGTGCTGACGGGAGTTGAACCCGCAATC

Glu_03from *Firmicutes bacterium M10-2*, anticodon: CTC, Seq. ID: W131196120

Seq.: GGCGCGTTCGGCAAGCGGTTAAGCCACAGCCCTCTCAAGGCTGTATCACGGGTTCAAATCCCGTACGCGCTGCCA

tREX probe:aaaaaaaaaaaaTGGCAGCGCGTACGGGATTTGAACCCGTGAT

Glu_04from *Enterococcus avium*, anticodon: CTC, Seq. ID: W131010917

Seq.: AACCCGTTAGTCAAGGGGTCAAGACGCCGCGTTCTCAGCGCGGAGGCACGGGTTCAAATCCCGTACGGGTTACCA

tREX probe:aaaaaaaaaaaaTGGTAACCCGTACGGGATTTGAACCCGTGCC

Glu_05from *Acaryochloris marina*, anticodon: TTC, Seq. ID: C08000515

Seq.: AGTGGTGTACGCAAATGGTTTAAGCGACTTGATTTTCAACCAAGTGATTGCGGGTTCAAATCCCGTCACCACTCCCA

tREX probe:aaaaaaaaaaaaTGGGAGTGGTGACGGGATTTGAACCCGCAAT

Glu_06from *Sporolactobacillus inulinus*, anticodon: CTC, Seq. ID: W11174757

Seq.: AGGGCGTTGGGCAAGTGGTTAAGCCAATGGATTCTCATTCCATGATCGCGGGTTCAATTCCCGTACGCCCTCCCA

tREX probe:aaaaaaaaaaaaTGGGAGGGCGTACGGGAATTGAACCCGCGAT

Glu_07from *Streptomyces davawensis*, anticodon: CTC, Seq. ID: C131020273

Seq.: CCGGATGTGGAGCAGAGGCAGACTCGCCGCCTTCTCAGGGCGGATACGCCGGTTCGAATCCGGCCGTCCGGGCCA

tREX probe:aaaaaaaaaaaaTGGCCCGGACGGCCGGATTCGAACCGGCGTA

Glu_08from *Leptolyngbya boryana*, anticodon: TTC, Seq. ID: W131036312

Seq.: GGTGCTGTAGGCAAATTGGTAAAGCCACAACTCTTTCAAAGTTGCGTTTGCGGGTTCAACTCCCGTCAGCACTGCCA

tREX probe:aaaaaaaaaaaaaTGGCAGTGCTGACGGGAGTTGAACCCGCAAAC

Glu_09from *Lactobacillus fructivorans*, anticodon: CTC, Seq. ID: W11144711

Seq.: TGCCCGTTGGTCAAACTGGTTTAAGACGTCGCCCTCTCAAGGCGGAGCTATGAGTTCAAGTCTCATACGGGTAACCA

tREX probe:aaaaaaaaaaaaTGGTTACCCGTATGAGACTTGAACTCATAGC

Glu_10from *Gordonia araii*, anticodon: TTC, Seq. ID: W121123713

Seq.: AGTCCACTAGCTCAGTTGGCAGAGCCATTGACTTTCCATCAATCGGTCGCCGGTTCGAGTCCGGCGTGGCGTACCA

tREX probe:aaaaaaaaaaaaaTGGTACGCCACGCCGGACTCGAACCGGCGACC

Gly_01from *Kitasatospora phosalacinea*, anticodon: GCC, Seq. ID: W141756150

Seq.:
GGAGAGTCGGTCGAGTGGCAAGGCAGCGGCTTGCCAAGCCGTGGTCGGGTGAAAGCCCGCGCGCGTTCGATCCGCGCACTCTCCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGGAGAGTGCGCGGATCGAACGCGCGCGGGCTTTCACCCGAC

Gly_02from *Dactylosporangium aurantiacum*, anticodon: GCC, Seq. ID: W141758047

Seq.: TGGGCCGTAGCGCAGAGGTCGACGCACCGCTTTGCCAGAGCGGAGGACGCCGGTTCGATTCCGGCCGGTCCACCCA

tREX probe:aaaaaaaaaaaaaTGGGTGGACCGGCCGGAATCGAACCGGCGTCC

Gly_03from *Microscilla marina*, anticodon: GCC, Seq. ID: w022708

Seq.: AGAAGGTTGGTGTAACTGGAAACACATTCGCCTGCCAAGCGAAAAATTGCAGGTTCGAGTCCTGTGCCTTCTTCCA

tREX probe:aaaaaaaaaaaaaTGGAAGAAGGCACAGGACTCGAACCTGCAATT

Gly_04from *Bacteroides vulgatus*, anticodon: CCC, Seq. ID: C004835

Seq.: GGACGATTAGCTCAGAGGCAGAGCATCAGCTTCCCAAGCTGAGGGTCGCGGGTTCAAGTCCCGTATCGTCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGACGATACGGGACTTGAACCCGCGACC

Gly_05from *Calothrix sp. 336/3*, anticodon: TCC, Seq. ID: C151081784

Seq.: CAAAAGTTAGCTCATTTGGTAGAGCAATCGACTTCCAATCGATCTGTAACAGGTTCAAATCCTGTACTTTTAACCA

tREX probe:aaaaaaaaaaaaaTGGTTAAAAGTACAGGATTTGAACCTGTTACA

Gly_06from *Microcoleus sp. PCC 7113*, anticodon: TCC, Seq. ID: C131004866

Seq.: TGGTGGATAGCTCAATGGGTAGAGCGCATGACTTCCAATCATGAAGTTGTTGGTTCGAGTCCAACTCCATCAGCCA

tREX probe:aaaaaaaaaaaaaTGGCTGATGGAGTTGGACTCGAACCAACAACT

Gly_07from *Deinococcus sp. 2009*, anticodon: GCC, Seq. ID: W131213920

Seq.: CGTCCGGTAGCTCAGGGGAAGAGCACCTCACTGCCAGTGAGGAGGCCAGCGGTTCGATTCCGTTCCGGACGTCCA

tREX probe:aaaaaaaaaaaaaTGGACGTCCGGAACGGAATCGAACCGCTGGCC

Gly_08from *Methanococcoides burtonii*, anticodon: GCC, Seq. ID: At0985

Seq.: ACACCAGTAGTGTAGCGGTTATCACCGGGCGTTGCCAACGCTCGAACTCGGGTTCGATTCCCGACTGGTGTACCA

tREX probe:aaaaaaaaaaaaTGGTACACCAGTCGGGAATCGAACCCGAGTT

Gly_09from *Staphylococcus aureus*, anticodon: GCC, Seq. ID: W141505172

Seq.: GGTCTCGTAGTGTAGCGGTTAACACGCCTGCCTGCCACGCAGGAGATCGCGGGTTCGATTCCCGTCGAGACCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGGTCTCGACGGGAATCGAACCCGCGATC

Gly_10from *Spiroplasma chrysopicola*, anticodon: ACC, Seq. ID: C131014728

Seq.: GATCAATTGATGGAATTGGTAGACATAGCTGATTACCAATCAGCTGTTGAAAAACGTGCGGGTTCAAGTCCCGTATTGATCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGATCAATACGGGACTTGAACCCGCACGTTTTTCAAC

His_01from *Kitasatospora azatica*, anticodon: GTG, Seq. ID: W141820006

Seq.: TGCCGGAAGGTGTGACTGGCAGCGCGCCTGGCTGTGCATCAGGCGATCGGGGTTCGATTCCTCCTCCGGCAGCCA

tREX probe:aaaaaaaaaaaaTGGCTGCCGGAGGAGGAATCGAACCCCGATC

His_02from *Streptomyces sp. NTK 937*, anticodon: GTG, Seq. ID: W141622045

Seq.: CCTTTCGTAGCTCAATGGCAGAGCCGCGCGTTGTGGTCGCGCTGATCCCGGTTCGACTCCGGGCGGAGGGACCA

tREX probe:aaaaaaaaaaaaTGGTCCCTCCGCCCGGAGTCGAACCGGGATC

His_03from *Candidatus Pelagibacter ubique*, anticodon: GTG, Seq. ID: C018211

Seq.: GCCTGCGTAGTATAACGGTTAGTACGATAGTTTGTGGAACTATAGGATTTGGTTCGATTCCAAGCGCGGGTACCA

tREX probe:aaaaaaaaaaaaTGGTACCCGCGCTTGGAATCGAACCAAATCC

His_04from *Afifella pfennigii*, anticodon: GTG, Seq. ID: W141562735

Seq.: GCGAACGTAGCTCAGTTGGTTAGAGCGTCGGATTGTGGCTCCGAAGGTCGGTGGTTCGAATCCACCCGTTCGTACCA

tREX probe:aaaaaaaaaaaaaaTGGTACGAACGGGTGGATTCGAACCACCGACC

His_05from *Ottowia sp. oral taxon 894*, anticodon: GTG, Seq. ID: C151090597

Seq.: TTGTTTGTGGCTCATTCGGTAGAGCAATGCACTGTGAATGCATCGGTATATCCGTAGCGGGTTCGAGTCCCGCCAAACAAACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTTTGTTTGGCGGGACTCGAACCCGCTACGGATATACC

His_06from *Mycobacterium virus Bongo*, anticodon: GTG, Seq. ID: PHG14101323

Seq.: CCTCTTGTAGCTCAATGGTAGAGCGCCGTTTTGTGAGACCGGTGACCTGCGTTCGATTCGCAGCAGGGGGACCA

tREX probe:aaaaaaaaaaaaaTGGTCCCCCTGCTGCGAATCGAACGCAGGTC

His_07from *Lyngbya confervoides*, anticodon: GTG, Seq. ID: W1511400808

Seq.: GCGAACGTAGCTCAGTTGGTTAGAGCGCTGGATTGTGGCTCCAGAGGTCGGTGGTTCAAATCCACCCGTTCGTACCA

tREX probe:aaaaaaaaaaaaaTGGTACGAACGGGTGGATTTGAACCACCGACC

His_08from *Bacillus cereus*, anticodon: GTG, Seq. ID: W131004730

Seq.: CAGTGTGTGGCGTAATGGTAACGCAGTGGACTGTGAATCCATGAATGAGAGTTCGATTCTCTCCATGCTGACCA

tREX probe:aaaaaaaaaaaaTGGTCAGCATGGAGAGAATCGAACTCTCATT

His_09from *Parcubacteria group bacterium GW2011_GWA2_40_23*, anticodon: GTG, Seq. ID: W1511604308

Seq.: GGCCGGGTAGCTCAATTGGTAGAGCAACAGACTGTGGATCTGTGTGTCGCGGGTTCGATGCCCGTCCCGGCTACCA

tREX probe:aaaaaaaaaaaaaTGGTAGCCGGGACGGGCATCGAACCCGCGACA

His_10from *Blastopirellula marina*, anticodon: ATG, Seq. ID: w013883

Seq.: AGACCGTTAGCTCAATGGTAGAGCACCAGACTATGAATCTGGCTGTAGCAAGTTCGAATCTTGCACGGTTCACCA

tREX probe:aaaaaaaaaaaaaTGGTGAACCGTGCAAGATTCGAACTTGCTACA

Ile_01from *[Clostridium] scindens*, anticodon: TAT, Seq. ID: W08003134

Seq.: CGGGATGTAGCGCAGATGGTAGAGCGACTGCCTTATATGCAGCAAGCCCCTGGTTCGAGTCCAGGCATCCCGACCA

tREX probe:aaaaaaaaaaaaaTGGTCGGGATGCCTGGACTCGAACCAGGGGCT

Ile_02from *Sinorhizobium fredii*, anticodon: AAT, Seq. ID: W121066177

Seq.: CGCTGGGTGGAGCAGCCCGGTAGCTCGTCAGGCTAATAACCTGAAGGCCGCAGGTTCAAATCCTGCCCCCGCAACCA

tREX probe:aaaaaaaaaaaaaTGGTTGCGGGGGCAGGATTTGAACCTGCGGCC

Ile_03from *Blautia hydrogenotrophica*, anticodon: TAT, Seq. ID: W09117204

Seq.: CGGAACATAGCGCAGTTGGTAGAGCACCTGTCTTATACACAGCAAGTCCCTAGTTCGATTCTAGGTGTTCCGACCA

tREX probe:aaaaaaaaaaaaaTGGTCGGAACACCTAGAATCGAACTAGGGACT

Ile_04from *Cyanothece sp. PCC 7424*, anticodon: TAT, Seq. ID: C09300063

Seq.: CGGAAGTTAGCTCATTTGGTAGAGCGATGGGTGTATCCCCCATAGGGAATAGGTTCAAATCCTATACTTTCAACCA

tREX probe:aaaaaaaaaaaaaTGGTTGAAAGTATAGGATTTGAACCTATTCCC

Ile_05from *Clostridiales bacterium VE202-21*, anticodon: TAT, Seq. ID: W131238843

Seq.: TGGGACATAGCTCAGAAGGTAGAGCACCTGGCTTATATCCAGCGTGTCACCGGTTCGAACCCGGTTGTCCCTACCA

tREX probe:aaaaaaaaaaaaaTGGTAGGGACAACCGGGTTCGAACCGGTGACA

Ile_06from *Salinispora arenicola*, anticodon: GAT, Seq. ID: W131169380

Seq.: GTCGTCGTGGCCAAGAAGGTGAAGGCACCGAGCTGATACCTCGGAGATGCGGTGGTTCGAACCCACCCGACGACACCA

tREX probe:aaaaaaaaaaaaaTGGTGTCGTCGGGTGGGTTCGAACCACCGCATC

Ile_07from *Rhizobium sp. CF080*, anticodon: GAT, Seq. ID: W121086610

Seq.:
GGTAGAGTGGCTGAGTGGTTGAAGGCATCGAACTGATACTTCGACGGGGAGTCGCTTCTTCCGAGGGTTCGAATCCCTCCTCTACCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGGTAGAGGAGGGATTCGAACCCTCGGAAGAAGCGACTCCCC

Ile_08from *Butyrivibrio sp. LC3010*, anticodon: AAT, Seq. ID: W131233184

Seq.: TTCAGAATGGCGAAGTGGAAATCGCGCTGGGTTAATGGCCCGGAAGTCGCAGGTTCGAATCCTGCTTCTGAATCCA

tREX probe:aaaaaaaaaaaaaTGGATTCAGAAGCAGGATTCGAACCTGCGACT

Ile_09from *Mycobacterium abscessus*, anticodon: GAT, Seq. ID: W121060940

Seq.: CGCATTGTGGCGCAGTTGGTTAGCGCGCCGACCTGATAAGTCGGAGGCCGCAGGTTCAATCCCTGCCGATGCGACCA

tREX probe:aaaaaaaaaaaaaTGGTCGCATCGGCAGGGATTGAACCTGCGGCC

Ile_10from *Chitinophaga pinensis*, anticodon: TAT, Seq. ID: C10103715

Seq.: TGATCTGTAGCTCAACGGTTAGAGCAACTGCCTTATACGCAGCAGGCTACCGGTTCAAATCCGGTCAGATCAACCA

tREX probe:aaaaaaaaaaaaaTGGTTGATCTGACCGGATTTGAACCGGTAGCC

Lys_01from *Salinispora arenicola*, anticodon: CTT, Seq. ID: C08008478

Seq.: CCCGCCGTAGCTCAGTGGTAGAGCACCCGGCTCTTAACCGGGAGGACGTTGGTTCGAGCCCAGCCGGCGAGCCCA

tREX probe:aaaaaaaaaaaaaTGGGCTCGCCGGCTGGGCTCGAACCAACGTCC

Lys_02from *Methanosarcina acetivorans*, anticodon: CTT, Seq. ID: At1403

Seq.: GGGCCCGTAGCTTAGTCAGGCAGAGCGATGGACTCTTAATCCATAGGCCGGGGGTTCAAATCCCTTCGGGCCCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGGGCCCGAAGGGATTTGAACCCCCGGCC

Lys_03from *Pyrobaculum aerophilum*, anticodon: CTT, Seq. ID: At0204

Seq.: GGGCCCGTAGCTCAGCCTGGTAGAGCGGCGGGCTCTTAACCCGTAGGTCGTGGGTTCGAATCCCACCGGGCCCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGGGCCCGGTGGGATTCGAACCCACGACC

Lys_04from *Sulfolobus acidocaldarius*, anticodon: CTT, Seq. ID: At0455

Seq.: GGGCCCGTAGCTCAGCCAGGTAGAGCGGCGGGCTCTTAACCCGTAGGTCCCGGGTTCAAATCCCGGCGGGCCCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGGGCCCGCCGGGATTTGAACCCGGGACC

Lys_05from *Methanoculleus marisnigri*, anticodon: CTT, Seq. ID: At1273

Seq.: GGGCCTGTAGCTTAGTTGGCAGAGCGACGGACTCTTAATCCGTAGGCCAAGGGTTCAAATCCCTTCAGGCCCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGGGCCTGAAGGGATTTGAACCCTTGGCC

Lys_06from *Sulfolobus solfataricus*, anticodon: CTT, Seq. ID: At0490

Seq.: GGGCCCGTAGCTTAGCCAGGTAGAGCGACGGGCTCTTAACCCGTAGTCCCGGGTTCGAATCCCGGCGGGCCCGCCA

tREX probe:aaaaaaaaaaaaTGGCGGGCCCGCCGGGATTCGAACCCGGGAC

Lys_07from *Microscilla marina*, anticodon: CTT, Seq. ID: w000994

Seq.: GTTGAGGTAGCTCAGTGGTAGAGCTTTCGGCGCTTAACCGAAAGGACGCAGGTTCGAGCCCTGCCCTCAATGCCA

tREX probe:aaaaaaaaaaaaaTGGCATTGAGGGCAGGGCTCGAACCTGCGTCC

Lys_08from *Candidatus Saccharimonas aalborgensis*, anticodon: CTT, Seq. ID: C131016148

Seq.: GGGCCAGTAGCTCAGCTGGTTAGAGCACCTGCCTCTTAAGCAGGGTGTCGAGAGTTCAAGTCTCTCCTGGCCCTCCA

tREX probe:aaaaaaaaaaaaaTGGAGGGCCAGGAGAGACTTGAACTCTCGACA

Lys_09from *Methanospirillum hungatei*, anticodon: CTT, Seq. ID: At1600

Seq.: GGGTCTGTAGCTTAGTCGGTAGAGCGGCGGACTCTTAATCCGCAGGCCAGGGGTTCGAGTCCCTTCAGGCCCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGGGCCTGAAGGGACTCGAACCCCTGGCC

Met_01from *Candidatus Koribacter versatilis*, anticodon: CAT, Seq. ID: C000175

Seq.: GGCGGCGTAGCTCAGCTGGTTAGAGCGACGGTCTCATAATCCGTAGGTCCGTGGTTCGAGTCCACGCGCCGCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGCGGCGCGTGGACTCGAACCACGGACC

Met_02from *Acidiphilium cryptum*, anticodon: CAT, Seq. ID: C000488

Seq.: GGCGGCGTAGCTCAGCGGTAGAGCAGGGGAATCATAATCCCTTGGTCGGTGGTTCAAATCCGCCCGCCGCTACCA

tREX probe:aaaaaaaaaaaaaTGGTAGCGGCGGGCGGATTTGAACCACCGACC

Met_03from *Aquifex aeolicus*, anticodon: CAT, Seq. ID: C000005

Seq.: GGCGGCGTAGCTCAGCTGGTCAGAGCGGGGATCTCATAAGTCCCAGGTCGGAGGTTCGAGTCCTCCCGCCGCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGCGGCGGGAGGACTCGAACCTCCGACC

Met_04from *Bacteroides fragilis*, anticodon: CAT, Seq. ID: C002823

Seq.: GGCGGGATAGCTCAGCTGGTTAGAGCGCATGATTCATAATCATGAGGTCCCCGGTTCAATCCCGGGTCCCGCTACCA

tREX probe:aaaaaaaaaaaaaTGGTAGCGGGACCCGGGATTGAACCGGGGACC

Met_05from *Acidovorax sp. JS42*, anticodon: CAT, Seq. ID: C000375

Seq.: GGTGGTATAGCTCAGTTGGTTAGAGCGCAGCATTCATAATGCTGATGTCCCAGGTTCAAGTCCCGGTACCACCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGTGGTACCGGGACTTGAACCTGGGACA

Met_06from *Chlamydia abortus*, anticodon: CAT, Seq. ID: C004944

Seq.: GGCGGTATAGCTCAGATGGTTAGAGCAGCAGAATCATAATCTGCGAGTCGTTGGTTCAAGTCCGACTACCGCTACCA

tREX probe:aaaaaaaaaaaaaTGGTAGCGGTAGTCGGACTTGAACCAACGACT

Met_07from *Dehalococcoides mccartyi*, anticodon: CAT, Seq. ID: C007625

Seq.: GGCAGCGTAGCTCAGTGGCAGAGCAGGGGACTCATAAGCCCTTGGTCGGTAGTTCAAATCTACCCGCTGCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGCAGCGGGTAGATTTGAACTACCGACC

Met_08from *Cenarchaeum symbiosum*, anticodon: CAT, Seq. ID: At0055

Seq.: GCCGGGATAGCTCAGCCGGTCAGAGTGCCAGACTCATAATCTGGAGGTCGTGGGATCAAAACCCACTCCCGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCGGGAGTGGGTTTTGATCCCACGACC

Met_09from *Nostoc punctiforme*, anticodon: CAT, Seq. ID: C08006893

Seq.: GTTTGATTAGCTCAGTTGGTAGAGCGATCGACTCATAATCGAAGGGTCACAGGTTCGACTCCTGTATCAAACCCCA

tREX probe:aaaaaaaaaaaaaTGGGGTTTGATACAGGAGTCGAACCTGTGACC

Met_10from *Sulfurimonas denitrificans*, anticodon: CAT, Seq. ID: C025436

Seq.: GTCAGGGTAGCTCAGCTGGTTAGAGCACTGGTCTCATAAGCCGGGGGTCGGGGGTTCGAGTCCCCCCTCTGACACCA

tREX probe:aaaaaaaaaaaaaTGGTGTCAGAGGGGGGACTCGAACCCCCGACC

Phe_01from *Bifidobacterium animalis*, anticodon: GAA, Seq. ID: C09300022

Seq.: GGCTCTGTAGCTCAGTTGGTAGAGCGAACGACTGAAAATCGTTAGGTCAGCGGATCGACGCCGCTCGGAGCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGCTCCGAGCGGCGTCGATCCGCTGACC

Phe_02from *Kocuria rhizophila*, anticodon: GAA, Seq. ID: C08012591

Seq.: GGCTCTGTAGCTCAGTTGGTAGAGCGTACGACTGAAAATCGTAAGGTCACCGGATCGACGCCGGTCGGAGCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGCTCCGACCGGCGTCGATCCGGTGACC

Phe_03from *Mobiluncus curtisii*, anticodon: GAA, Seq. ID: C10300161

Seq.: GGCTCTGTAGCTCAGTTGGCAGAGCGAACGACTGAAAATCGTTAGGTCACCGGATCGACGCCGGTCGGAGCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGCTCCGACCGGCGTCGATCCGGTGACC

Phe_04from *Candidatus Aquiluna sp. IMCC13023*, anticodon: GAA, Seq. ID: W123001457

Seq.: GGCTCTGTAGCTCAGTTGGTAGAGCGAACGACTGAAAATCGTTAGGTCACCGGATCGACGCCGGTCGGAGCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGCTCCGACCGGCGTCGATCCGGTGACC

Phe_05from *Picrophilus torridus*, anticodon: GAA, Seq. ID: At1756

Seq.: GCCGTGATAGCTCAGTTGGGAGAGCGCCAGACTGAAGATCTGGAGGTCGCTGGTTCAATCCCGGCTCACGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCGTGAGCCGGGATTGAACCAGCGACC

Phe_06from *Aeropyrum pernix*, anticodon: GAA, Seq. ID: At0036

Seq.: GCCGCCGTAGCTCAGCGGGAGAGCGCCCGGCTGAAGACCGGGTGGTCCGGGGTTCGAATCCCCGCGGCGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCGCCGCGGGGATTCGAACCCCGGACC

Phe_07from *Chitinophagaceae bacterium JGI 0001002-J12*, anticodon: GAA, Seq. ID: W131214335

Seq.: GGACGAATAGCTCCAATGGCAGAGCATCACGTTGAAGCCGTGACAGTGTTGGTTCGAATCCAGCTTCGTCTACCA

tREX probe:aaaaaaaaaaaaTGGTAGACGAAGCTGGATTCGAACCAACACT

Phe_08from *Actinomyces coleocanis*, anticodon: GAA, Seq. ID: W09300480

Seq.: GCCTCTGTAGCTCAGTTGGTAGAGCGTTCGACTGAAAATCGAAAGGTCACCGGATCGACGCCGGTCGGAGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCTCCGACCGGCGTCGATCCGGTGACC

Phe_09from *Microbacterium testaceum*, anticodon: GAA, Seq. ID: C11300378

Seq.: GGCTCTGTAGCTCAGTTGGTAGAGCGCACGACTGAAAATCGTGAGGTCACGGGATCGACGCCCGTCGGAGCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGCTCCGACGGGCGTCGATCCCGTGACC

Phe_10from *Ignicoccus hospitalis*, anticodon: GAA, Seq. ID: At2227

Seq.: GCCGCCGTAGCTCAGCGGGAGAGCGCCCGGCTGAAGACCGGGTCGTCCGGGGTTCGAATCCCCGCGGCGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCGCCGCGGGGATTCGAACCCCGGACG

Pro_01from *Nocardia otitidiscaviarum*, anticodon: TGG, Seq. ID: W141231342

Seq.: TTGGGTGTAGTTCAATTTGGTCAGAGCACCCGCTTTGGGAGCGGGAAGTTGCGAGTTCGAGTCTCGCCACCCGAACCA

tREX probe:aaaaaaaaaaaaaTGGTTCGGGTGGCGAGACTCGAACTCGCAACT

Pro_02from *Acetobacter pasteurianus*, anticodon: TGG, Seq. ID: W121126922

Seq.:
GCGGCCATGGTGAAATTTGGTAAACACATATAATTTGGAATTATAGATACATTTAAGAGAGTATTTGAGGGTTCAAGTCCCTCTGGCCGCACCA

tREX probe:aaaaaaaaaaaaaaaaaaaaTGGTGCGGCCAGAGGGACTTGAACCCTCAAATACTCTCTTAAATGTAT

Pro_03from *Aureispira sp. CCB-QB1*, anticodon: TGG, Seq. ID: W141526482

Seq.: ACGGGAGTAGCTTAACAAGGCAAAAGCACTAGGTTTGGGACTTAGGAGATGCAGGTTCAACCCCTGTCACCCGTACCA

tREX probe:aaaaaaaaaaaaaTGGTACGGGTGACAGGGGTTGAACCTGCATCT

Pro_04from *Streptococcus pyogenes*, anticodon: TGG, Seq. ID: C151019424

Seq.:
GGAGGATTACCCAAGTCCGGCTGAAGGGAACGGTCTTGGAAACCGACAAAGGAGTAAAAGAGTGGGTGGGTTCGAATCCCACATCCTCCTCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGAGGAGGATGTGGGATTCGAACCCACCCACTCTTTTACTCCTTT

Pro_05from *Beggiatoa alba*, anticodon: TGG, Seq. ID: W123000923

Seq.: GTCCAATTAGCTCACAGGTTAGAGCGTAATTACTGGGTAATTAAGGTAGCAGGTTCAACTCCTGCATTGGACTCCA

tREX probe:aaaaaaaaaaaaaTGGAGTCCAATGCAGGAGTTGAACCTGCTACC

Pro_07from *Parcubacteria group bacterium GW2011_GWA2_46_10*, anticodon: TGG, Seq. ID: W1511614445

Seq.: CGGGCTGTAGTGTCAACGGCCAGCACGCGTGCTTTGGGAGCACGTAGATCGAGTTCGAATCTCGACAGCCCGCCCA

tREX probe:aaaaaaaaaaaaTGGGCGGGCTGTCGAGATTCGAACTCGATCT

Pro_08from *Coprobacillus sp. D7*, anticodon: TGG, Seq. ID: W09119425

Seq.: TGAGAAGTAGCACAATTTGGTAGTGCTCGTGGTTTGGGACCACGGGGTTATGGGTTCAAATCCCATCTTCTCAACCA

tREX probe:aaaaaaaaaaaaaTGGTTGAGAAGATGGGATTTGAACCCATAACC

Pro_09from *Lachnospiraceae bacterium A4*, anticodon: TGG, Seq. ID: W131195543

Seq.: TACTTCGTAGCTCATCTGGTAGAGCAATTTGTTTGGGACAAATTTGTAGCTGGTTCGAATCCAGTCGAAGGCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGCCTTCGACTGGATTCGAACCAGCTACA

Pro_10from *Mucilaginibacter paludis*, anticodon: TGG, Seq. ID: W11138394

Seq.: AGGGTAATATATCAATTGGCAGATTACTTGCTTTGGGAGCATGAGGTTGTGGGTTCGAGTCCCGCTTACCCTACCA

tREX probe:aaaaaaaaaaaaaTGGTAGGGTAAGCGGGACTCGAACCCACAACC

Thr_01from *Amycolatopsis orientalis*, anticodon: CGT, Seq. ID: W131202873

Seq.: CATGGAGTAGATCAGCGGCAGATCGCCCGGCTCGTAACCGGGAGGGCGCGGGTTCGAGTCCCGCCTCCGTGTCCA

tREX probe:aaaaaaaaaaaaaTGGACACGGAGGCGGGACTCGAACCCGCGCCC

Thr_02from *Chlamydia trachomatis*, anticodon: CGT, Seq. ID: W1510645195

Seq.: TGCGGGGTAGTGTAACGGTAACATACCGGTTTCGTTTTCCGGTGCTGCGGGTTCGATTCCCGCCCCCGCCTCCA

tREX probe:aaaaaaaaaaaaTGGAGGCGGGGGCGGGAATCGAACCCGCAGC

Thr_03from *Lyngbya sp. PCC 8106*, anticodon: AGT, Seq. ID: w021635

Seq.: CGGTGGATAGCTCAATAGGTAGAGCGCATGATTAGTCATCATGAGTTGTTGGTTCGATTCCAACTCCATCAACCA

tREX probe:aaaaaaaaaaaaTGGTTGATGGAGTTGGAATCGAACCAACAAC

Thr_04from *Intestinimonas butyriciproducens*, anticodon: TGT, Seq. ID: W1511303119

Seq.: TGCGGGGTGGTGTAGTGGTGAACATTAGTCTCTTGTCATGACTGGCCGCTGGTTCGATTCCAGCCCCCGCACCCA

tREX probe:aaaaaaaaaaaaTGGGTGCGGGGGCTGGAATCGAACCAGCGGC

Thr_05from *Lachnospiraceae bacterium AC2014*, anticodon: TGT, Seq. ID: W141731478

Seq.: TACGAAGTGGCGCAATAGGTAGCGCTCATGATTTGTAATCGTGAAGTTGTAGGTTCGAGTCCTATCTTTGTAACCA

tREX probe:aaaaaaaaaaaaaTGGTTACAAAGATAGGACTCGAACCTACAACT

Thr_06from *Candidatus Jorgensenbacteria bacterium GW2011_GWA1_48_11*, anticodon: CGT, Seq. ID: W1511617203

Seq.: CGCCGGGTGGAGCAGCCTGGTAGCTCGTCGGGCTCGTAGCCCGAAGGTCGCAGGTTCAAATCCTGCCCCGGCTACCA

tREX probe:aaaaaaaaaaaaaTGGTAGCCGGGGCAGGATTTGAACCTGCGACC

Thr_07from *Enterococcus faecalis*, anticodon: CGT, Seq. ID: W10125366

Seq.: TGGCGTGTAGCATTGTGGTAATGCAACTTACTTCGTGTGAGATAAGATGCGGGTTCGAATCCTGTCACGCCAACCA

tREX probe:aaaaaaaaaaaaaTGGTTGGCGTGACAGGATTCGAACCCGCATCT

Thr_08from *Lactobacillus reuteri*, anticodon: GGT, Seq. ID: W11186032

Seq.: AACACGTTAGTTTAAATGGGGAAGCACCACTATGGTAATGTGGAGATATGGGTTCGAATCCTGTACGTGTTACCA

tREX probe:aaaaaaaaaaaaTGGTAACACGTACAGGATTCGAACCCATATC

Thr_09from *Alistipes putredinis*, anticodon: AGT, Seq. ID: W08002770

Seq.: TGCGGGGTAGTGCAGAGGTTACCACGGCGGGTTAGTGTCCCGCAGGCGCAAGTTCGATTCTTGCCCCCGCTACCA

tREX probe:aaaaaaaaaaaaTGGTAGCGGGGGCAAGAATCGAACTTGCGCC

Thr_10from *Calothrix sp. PCC 7507*, anticodon: TGT, Seq. ID: C131010288

Seq.: AGGAAGTTAGCTCACTGGTAGAGCGATCGACTTGTAATCGATTTGTTATAGGTTCGATTCCTATACTTTCAACCA

tREX probe:aaaaaaaaaaaaaTGGTTGAAAGTATAGGAATCGAACCTATAACA

Trp_01from *Leucobacter musarum*, anticodon: CCA, Seq. ID: W1511008762

Seq.: GTCTGCGTAGCTCAATTGGCTAGAGCACCGGTCTCCAAAACCGGAGGTTCCAGGTTCGAGTCCTGGCGCATTCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGAATGCGCCAGGACTCGAACCTGGAACC

Trp_02from *Ensifer sojae*, anticodon: CCA, Seq. ID: W121066370

Seq.: TACCGGTTAGCTTAGTGGTAAAGCGTCGGTCTCCAAAACCGATTACGTGGGTTCGATTCCCGCACCGGGTGCCA

tREX probe:aaaaaaaaaaaaTGGCACCCGGTGCGGGAATCGAACCCACGTA

Trp_03from *Nonlabens sediminis*, anticodon: CCA, Seq. ID: W141238833

Seq.: GCAGGTTTAGCTCAGTTGGTAGAGCACTGGTCTCCAAAACCAGGTGTCGGGAGTTCGAGCCTCTCAACCTGCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGCAGGTTGAGAGGCTCGAACTCCCGACA

Trp_04from *Chryseobacterium sp. UNC8MFCol*, anticodon: CCA, Seq. ID: W141168454

Seq.: ACTCCGATAGTGTAACGGCAGTATCTCAGTCTCCAAAACTGACAGTACTGGTTCGAATCCGGTTCGGAGTTCCA

tREX probe:aaaaaaaaaaaaTGGAACTCCGAACCGGATTCGAACCAGTACT

Trp_05from *Achromobacter sp. DH1f*, anticodon: CCA, Seq. ID: W141187535

Seq.: TCCGGGATAGTTCAGCGGTAGAACATCCGGCTCCACTCTGGATAGGCGCTGGTTCGACTCCAGCTCCCGGATCCA

tREX probe:aaaaaaaaaaaaaTGGATCCGGGAGCTGGAGTCGAACCAGCGCCT

Trp_06from *Caulobacter virus Swift*, anticodon: CCA, Seq. ID: PHG14101650

Seq.: GTCGGTCTAGCTCATGGGGTAGAGCGGCGGTCTCCAAAACCGCGCGTGGCAGGTTCGAGTCCTGCGACCGGCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGCCGGTCGCAGGACTCGAACCTGCCACG

Trp_07from *Deinococcus misasensis*, anticodon: CCA, Seq. ID: W141816918

Seq.: GATCTGGTAGCTCAGTCTGGTAGAGCGTGCGTCTCCAAAACGCAGGGTCGCAGGTTCAAATCCTGCCCGGATCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGATCCGGGCAGGATTTGAACCTGCGACC

Trp_08from *Vibrio parahaemolyticus*, anticodon: CCA, Seq. ID: W1511591459

Seq.: GATTTCGTGGCGCAATGGTAGCGCGCCTGACTCCAGATCAGGAGGTTGCGTGTTCAAATCACGTCGAGATCACCA

tREX probe:aaaaaaaaaaaaaTGGTGATCTCGACGTGATTTGAACACGCAACC

Trp_09from *Mycoplasma hyorhinis*, anticodon: TCA, Seq. ID: C11114494

Seq.: tAGAGGTGTAGTTCAATGGTAGAACAACGGGCTTCAACCCCGTGTGTTGCGGGTTCGACTCCTGTCACCTCTGCCCA

tREX probe:aaaaaaaaaaaaaTGGGCAGAGGTGACAGGAGTCGAACCCGCAACA

Trp_10from *Mycoplasma putrefaciens*, anticodon: TCA, Seq. ID: C11115064

Seq.: cAGGGGCATAGTTCAGTTGGTAGAACATCGGTCTTCAAAACCGAGTGTCACGAGTTCGAGTCTTGTTGCCCCTGCCCA

tREX probe:aaaaaaaaaaaaaTGGGCAGGGGCAACAAGACTCGAACTCGTGACA

Tyr_01from *Mycobacterium abscessus*, anticodon: GTA, Seq. ID: W123001397

Seq.: CGTGGGTTGGGCTTGTTGGGCAAGGCCCTCCAGACTGTAAATCTGGCGCTTCGGCTACGGGGGTTCGATTCCCTCCCTACGTCCA

tREX probe:aaaaaaaaaaaaaaaTGGACGTAGGGAGGGAATCGAACCCCCGTAGCCGAAGC

Tyr_02from *Rhodobacter sphaeroides*, anticodon: GTA, Seq. ID: C027996

Seq.: TGGGCGACGTGCTGCAAGGTGCGGCAGCGGACTGTAACTCCGCCGAGGCGACTCGTGCCTGGTTCGATTCCAGGGTCGCCCACCCA

tREX probe:aaaaaaaaaaaaaaaaaTGGGTGGGCGACCCTGGAATCGAACCAGGCACGAGTCGCCTC

Tyr_03from *Pseudopedobacter saltans*, anticodon: ATA, Seq. ID: W11140046

Seq.:
TGTAGGGTGGTGAAATGGCAGACACACCTCCTTATAATGGTGAAGAAAGTTCAAAAGACTGGATTCTTGTAGGTTCGACTCCTACCCCTACAGCCA

tREX probe:aaaaaaaaaaaaaaaaaaaaaTGGCTGTAGGGGTAGGAGTCGAACCTACAAGAATCCAGTCTTTTGAACTTTC

Tyr_04from *Mycobacterium abscessus*, anticodon: GTA, Seq. ID: W122000209

Seq.: CTGCGCGCGGTGACGACGGCGTGTCACGACTGACTGTAAATCAGTTGCCTTTCGGCTCAGGGGGTTCAAATCCCTCCGCGCAGACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTCTGCGCGGAGGGATTTGAACCCCCTGAGCCGAAAGGC

Tyr_05from *Archaeoglobus fulgidus*, anticodon: GTA, Seq. ID: At0643

Seq.: CCCGCCTTAGCTCAGAGGTAGAGCGTGGGACTGTAGATCCCATGGTCCCCGGTTCAAATCCGGGAGGCGGGACCA

tREX probe:aaaaaaaaaaaaaTGGTCCCGCCTCCCGGATTTGAACCGGGGACC

Tyr_06from *Hungatella hathewayi*, anticodon: GTA, Seq. ID: W09125611

Seq.: GGATGCGTAAGCCCAGCGGCGAGGGCAGGAGACTGTAAATCTCTCACATCGGAAACAACGCAGGTTCGAGTCCTGCCGCATCCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGGATGCGGCAGGACTCGAACCTGCGTTGTTTCCGATG

Tyr_07from *Vibrio parahaemolyticus*, anticodon: GTA, Seq. ID: W1511591454

Seq.:
CCTTCGATAGCTCAGTTGGTAGAGCGGTGGACTGTAGCATTGTACAAAGATATCCATAGGTCACTGGTTCAACTCCGGTTCGAAGGACCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGTCCTTCGAACCGGAGTTGAACCAGTGACCTATGGATATCTTTG

Tyr_08from *Mycobacterium phage Catera*, anticodon: GTA, Seq. ID: PHG0200041

Seq.: CCCGTACATGCCCAACTGGTGTTTGGGAGCAGGCTGTAAATCTGTGGCCTTCGGGACGGTGAGGTTCGATTCCTCAGTGCGGGACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTCCCGCACTGAGGAATCGAACCTCACCGTCCCGAAGGC

Tyr_09from *Nanoarchaeum equitans*, anticodon: GTA, Seq. ID: At2113

Seq.: CCGGGCGTAGCTCAGCGGCAGAGCGGCCGGCTGTAGACCGGCAGGTCGGGGGTTCGAATCCCCCCGCCCGGACCA

tREX probe:aaaaaaaaaaaaaTGGTCCGGGCGGGGGGATTCGAACCCCCGACC

Tyr_10from *Parcubacteria group bacterium GW2011_GWA2_42_14*, anticodon: GTA, Seq. ID: W1511607542

Seq.: AGGGAGGTGGCCGAGTCCGGTTGAAGGCGCCAGACTGTAAATCTGGTGTGAAAGCCGCGTAGGTTCAAATCCTACCCTCCCTACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTAGGGAGGGTAGGATTTGAACCTACGCGGCTTTCAC

Val_01from *Paenibacillus dendritiformis*, anticodon: GAC, Seq. ID: W121034225

Seq.: GTTCTGATAGCTCAGTAGGGAGAGCACCATCTTGACAGGGTGGGGGTCGGCGGTTCGAGCCCGTCTCAGAACACCA

tREX probe:aaaaaaaaaaaaaTGGTGTTCTGAGACGGGCTCGAACCGCCGACC

Val_02from *Acetobacterium woodii*, anticodon: TAC, Seq. ID: C121003905

Seq.: GGTCGCATAGCTCAGCTGGGAGAGCACCTGCCTTACAAGCAGGGGGTCACAGGTTCGAGCCCTGTTGCGACCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGTCGCAACAGGGCTCGAACCTGTGACC

Val_03from *Enterococcus faecalis*, anticodon: TAC, Seq. ID: W133000314

Seq.: GGAGGTTTAGCTCAGCTGGGAGAGCACCTGCCTTACAAGCAGGGGGTCAGCGGTTCGATCCCGTTAACCTTCTCCA

tREX probe:aaaaaaaaaaaaaTGGAGAAGGTTAACGGGATCGAACCGCTGACC

Val_04from *Helicobacter pylori*, anticodon: GAC, Seq. ID: W123001937

Seq.: GGTCGCGTAGCTCAGTTGGTAGAGCACTACCTTGACATGGTAGTGGCCGCTGGTTCAAGTCCAGTCGTGGCGCCCA

tREX probe:aaaaaaaaaaaaaTGGGCGCCACGACTGGACTTGAACCAGCGGCC

Val_05from *Mycoplasma anatis*, anticodon: TAC, Seq. ID: W11174739

Seq.: GGAAGATTAGCTCAGTTGGGAGAGCGTTGCCCTTACAAGGCAAATGTCATGGGTTCGAGTCCCTTATCTTCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGAAGATAAGGGACTCGAACCCATGACA

Val_06from *Campylobacter showae*, anticodon: GAC, Seq. ID: W131099323

Seq.: GGTCCCGTAGCTCAGTTGGTAGAGTACTACCTTGACATGGTAGTGGTCGATGGTTCGAGTCCATTCGGGGCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGCCCCGAATGGACTCGAACCATCGACC

Val_07from *Candidatus Mycoplasma haemolamae*, anticodon: TAC, Seq. ID: C121015518

Seq.: GGAATGTTAGCTCAGCGGGAGAGCAACTGCCTTACAAGCAGTAGGTCGGGGGTTCAAATCCCTCACATTCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGAATGTGAGGGATTTGAACCCCCGACC

Val_08from *Mycoplasma cynos*, anticodon: TAC, Seq. ID: C151099846

Seq.: GGAAGATTAGCTCAGTTGGGAGAGCGTCGCCCTTACAAGGCGAATGTCATGGGTTCGAGTCCCTTATCTTCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGAAGATAAGGGACTCGAACCCATGACA

Val_09from *Bacillus cereus*, anticodon: GAC, Seq. ID: W131005331

Seq.: GATCCCGTAGCTCAGCAGGGAGAGCGCCACCTTGACAGGGTGGAGGTCGTGAGTTCGAGCCTCACTGGGGTCACCA

tREX probe:aaaaaaaaaaaaaGGTGACCCCAGTGAGGCTCGAACTCACGACC

Val_10from *Butyrivibrio sp. VCD2006*, anticodon: TAC, Seq. ID: W131232029

Seq.: TCCGATATAACTCAGCTGGGAGAGTGCTTCCCTTACAAGGAAGAAGCCACAGGTTCAATCCCTGTTATCGGAACCA

tREX probe:aaaaaaaaaaaaaTGGTTCCGATAACAGGGATTGAACCTGTGGCT

Trp_11from *Actinosynnema mirum*, anticodon: CCA, Seq. ID: C09100630

Seq.: CGTGGAGTAGCTCAGGTGGTAGAGCTCCCGGCTCCAAACCGGGAAGTCGCGGGTTCGACTCCCGCCTCCATGTCCA

tREX probe:aaaaaaaaaaaaaTGGACATGGAGGCGGGAGTCGAACCCGCGACT

Trp_12from *Microcoleus vaginatus*, anticodon: CCA, Seq. ID: W11164215

Seq.: GCAGGGATGGTGTAACTGGCAACACTCTGGTCTCCAAAACCAGCATTCTGAGTTCAAATCTCAGTTCCTGTGCCA

tREX probe:aaaaaaaaaaaaTGGCACAGGAACTGAGATTTGAACTCAGAAT

Trp_13from *Desulfatibacillum alkenivorans*, anticodon: CCA, Seq. ID: C09103670

Seq.: GTGCCCTTGGCGTCAACTGGCAGCGCAACGGACTCCAAATCCGTAGGTTGAAGGTTCGAATCCTTCAGGGTACGCCA

tREX probe:aaaaaaaaaaaaaTGGCGTACCCTGAAGGATTCGAACCTTCAACC

Trp_14from *Faecalibacterium prausnitzii*, anticodon: CCA, Seq. ID: W11133128

Seq.: GCCGTTTTAGCTCATGTTGGAAGAGCGCCGGTCTCCAAAGCCGGAAGCGGCAGGTTCGATCCCTGCAAACGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCGTTTGCAGGGATCGAACCTGCCGCT

Trp_15from *Phycisphaera mikurensis*, anticodon: CCA, Seq. ID: C121001250

Seq.: GCCGGTGTAGCTCAGTTGGTTAGAGCAGTGGATTCCAAATCCACAGGTCGCGGGTTCGAGTCCCTCCGCCGGTGCCA

tREX probe:aaaaaaaaaaaaaTGGCACCGGCGGAGGGACTCGAACCCGCGACC

Trp_16from *Leptolyngbya boryana*, anticodon: CCA, Seq. ID: W131036311

Seq.: GTTGGGATGGTGTAACGGTAACATTTCGGTCTCCAAAACCGACGATCTGAGTTCGAGTCTCAGTCCCTTCGCCA

tREX probe:aaaaaaaaaaaaTGGCGAAGGGACTGAGACTCGAACTCAGATC

Trp_17from *Anabaena cylindrica*, anticodon: CCA, Seq. ID: C131005049

Seq.: GCACAGATGGTGTAACGGAAGCATAAGGGTCTCCAAAACCTTTGGTCATGGTTCAAATCCTTGTCTGTGCGCCA

tREX probe:aaaaaaaaaaaaTGGCGCACAGACAAGGATTTGAACCATGACC

Trp_18from *Acidobacteriaceae bacterium KBS 96*, anticodon: CCA, Seq. ID: W131179233

Seq.: GCCGGGATCGTTCAGCGGCTAGGACGGCTGCCTCCAGAACAGCTTACGAGGGTTCGAGTCCTTCTCCCGGTGCCA

tREX probe:aaaaaaaaaaaaTGGCACCGGGAGAAGGACTCGAACCCTCGT

Trp_19from *Clostridium botulinum*, anticodon: CCA, Seq. ID: W08001255

Seq.: GTAGGTATGGTGTAATGGCTAACATGATAGTCTCCAAAACTATTGATGTGGGTTCGATTCCTACTACCTATGCCA

tREX probe:aaaaaaaaaaaaTGGCATAGGTAGTAGGAATCGAACCCACATC

Trp_20from *Calothrix sp. PCC 7103*, anticodon: CCA, Seq. ID: W131036258

Seq.: GCAGCGATGGTGTAATTGGCTAACACGTCGGTCTCCAAAACCGAAGAATCTAGGTTCAAGCCCTAGTCGCTGCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGCAGCGACTAGGGCTTGAACCTAGATTC

Ser_2682from *Arcobacter butzleri*, anticodon: GCT, Seq. ID: C11100375

Seq.:
GGACAGTTGGGTGAGTTGGCTGAAACCACCTCCCTGCTAAGGAGACGTACTGGTAACGGTACCGAGGGTTCAAATCCCTCACTGTCCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGGACAGTGAGGGATTTGAACCCTCGGTACCGTTACCAGTACG

Ser_3918from *Mycoplasma hominis*, anticodon: GCT, Seq. ID: C121016261

Seq.:
GGGTAAGTACTCAAGTGGTCGAAGAGGCGGTCCTGCTAAGACTGTAGGGGTGCTAAAACACCCGCGGAGGTTCAAATCCTCTCTTACCCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaaaTGGCGGGTAAGAGAGGATTTGAACCTCCGCGGGTGTTTTAGCACCCCT

Ser_3137from *Candidatus Carsonella ruddii*, anticodon: TGA, Seq. ID: C121014325

Seq.:
GGAAAGATGGTAGAGTGGATTAATACGTTGGTCTTGAAAACCAAAAAAGTTTATACTTTCCAGGGTTCGAATCCCTGTCTTTCCACCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGTGGAAAGACAGGGATTCGAACCCTGGAAAGTATAAACTTT

Ser_10084from *Hymenobacter norwichensis*, anticodon: GCT, Seq. ID: W131215968

Seq.:
AGAGAGATGGGTGAGTGGCTTAAACCAGTAGTTTGCTAAACTGCCGTAGCTCTAAAGGTTACCGGGGGGTTCGAATCCCCCTCTCTCTACCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGTAGAGAGAGGGGGATTCGAACCCCCGGTAACCTTTAGAGCTAC

Ser_5064from *Cyclobacterium marinum*, anticodon: GCT, Seq. ID: C11105728

Seq.:
AGAGAGTTGGGTGAGTGGCTTAAACCAGCAGTTTGCTAAACTGTCGTACGCCTAATAGCGTACCGGGGGGTTCGAATCCCCCACTCTCTACCA

tREX probe:aaaaaaaaaaaaaaaaaaaaTGGTAGAGAGTGGGGGGATTCGAACCCCCGGTACGCTATTAGGCGTAC

Ser_7617from *Dyadobacter alkalitolerans*, anticodon: GCT, Seq. ID: W131221754

Seq.:
GGAGGGATGGGTGAGTGGCTGAAACCAGTAGTTTGCTAAACTGCCGTACTCGCAAGGGTACCGCGGGTTCGAATCCCGCTTCCTCTACCA

tREX probe:aaaaaaaaaaaaaaaaaaaTGGTAGAGGAAGCGGGATTCGAACCCGCGGTACCCTTGCGAGTAC

Ser_6664from *Echinicola pacifica*, anticodon: GCT, Seq. ID: W131167768

Seq.:
AGAGAGTTGGGTGAGTGGCTTAAACCAGCAGTTTGCTAAACTGTCGTACGCGTAATAGTGTACCGGGGGGTTCGAATCCCCCACTCTCTACCA

tREX probe:aaaaaaaaaaaaaaaaaaaaTGGTAGAGAGTGGGGGATTCGAACCCCCGGTACACTATTACGCGTAC

Ser_1282from *Lactococcus garvieae*, anticodon: GCT, Seq. ID: C11112727

Seq.:
GGGGAGTTACTCAAGAGGCTGAAGAGGACGGTTTGCTAAATCGTTAGGTCGGGAAACCGGCGCGAGGGTTCGAATCCCTTACTCCCCTCCA

tREX probe:aaaaaaaaaaaaaaaaaaaTGGAGGGGAGTAAGGGATTCGAACCCTCGCGCCGGTTTCCCGACCT

Ser_477from *Campylobacter hominis*, anticodon: TGA, Seq. ID: C08003181

Seq.:

CGGTAGATGGCTGAGCGGTCGAAAGCGGCGGTCTTGAAAACCGTTGAGGTGCAAGCCTCCTGGGGTTCGAATCCCTATCTACCGGCCA

tREX probe:aaaaaaaaaaaaaaaaaTGGCCGGTAGATAGGGATTCGAACCCCAGGAGGCTTGCACCTC

Ser_4565from *Streptobacillus moniliformis*, anticodon: GCT, Seq. ID: C10112044

Seq.:
GGATAAGTGGTAGAGAGGCCGAATACACTTCCCTGCTAAGGAAGCATCCGGGCATAAACCTGGATCGTGGGTTCAAATCCCACCTTGTCCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaaaTGGCGGACAAGGTGGGATTTGAACCCACGATCCAGGTTTATGCCCGGAT

Ser_3958from *Mesoplasma florum*, anticodon: GCT, Seq. ID: C013434

Seq.:
GGGTTAATACTCAAGTTGGTGAAGAGGACACCCTGCTAAGGTGTTAGGTCGGTTTCCGGCGCAAGAGTTCGAGTCTCTTTTAACCCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGGGTTAAAAGAGACTCGAACTCTTGCGCCGGAAACCGACCT

Ser_9037from *Mycoplasma bovigenitalium*, anticodon: CGA, Seq. ID: W131098181

Seq.:
CGGTAGTTGCTCGAGTGGCTGAAGAGGTTTGTCTCGAAAACAAATAGTCGCGCAAGCGGCTCAAGGGTTCAAATCCCTTACTACCGGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCCGGTAGTAAGGGATTTGAACCCTTGAGCCGCTTGCGCGACT

Ser_7016from *Methyloferula stellata*, anticodon: TGA, Seq. ID: W131186337

Seq.: GACCGGGTGGCAGAATGGCTATGCAGGGCTCTTGAAAATCTCGCACGCCGGTTCGATTCCGGCCCCGGTCTCCA

tREX probe:aaaaaaaaaaaaTGGAGACCGGGGCCGGAATCGAACCGGCGTG

Ser_3135from *Bulleidia extructa*, anticodon: GCT, Seq. ID: W10113111

Seq.:
GAGGAAATACTCAAGAGGCCGAAGAGGTGCCCCTGCTAAGGGCATAGGTCGAGAAATCGGCGCGAGGGTTCAAATCCCTCTTTCCTCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGAGGAAAGAGGGATTTGAACCCTCGCGCCGATTTCTCGACCT

Ser_6589from *Oribacterium sp. oral taxon 078*, anticodon: GGA, Seq. ID: W09125722

Seq.: GACGAGGTGTGATAATGGTAGTCGGCCAGCCTGGAAAGTTGGTGCCCGTTCTGCGGGTTGTGGGTTCAAGTCCCATCCTCGTCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGACGAGGATGGGACTTGAACCCACAACCCGCAGAACGGGC

Ser_4348from *Solobacterium moorei*, anticodon: GCT, Seq. ID: W11132983

Seq.:
GAGGAAATACTCAAGAGGCCGAAGAGGTGCCCCTGCTAAGGGCATAGGTCGGGAAACCGGCGCGAGGGTTCAAATCCCTCTTTCCTCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGAGGAAAGAGGGATTTGAACCCTCGCGCCGGTTTCCCGACCT

Ser_8994from *Pacificimonas flava*, anticodon: GCT, Seq. ID: W131045831

Seq.:
GGGCACGTGGGTGAGTGGCTGAAACCAGCGGTTTGCTAAACCGCCATACGGGGTTGACCCGTATCGAGGGTTCGAATCCCTCCGTGCCCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaaaTGGCGGGCACGGAGGGATTCGAACCCTCGATACGGGTCAACCCCGTAT

Ser_3838from *Chlamydia psittaci*, anticodon: CGA, Seq. ID: W131211446

Seq.:
CGATGGTTGCTCGAGTGGCTGAAGAGGTTAGTCTCGAAAACTAATAATGTTAATAGCATTCAAGGGTTCAAATCCCTTACCATCGGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCCGATGGTAAGGGATTTGAACCCTTGAATGCTATTAACATT

Ser_4103from *Xanthomonas fuscans*, anticodon: TGA, Seq. ID: W10200001

Seq.: GGGTTGTTCCGCTGCGTTGGGCGGCGAGCGGTCTTGAAAACCGTGGTGGCCGAAAGGCCAGGGGTTCGACTCCTCAACAACCCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGGGTTGTTGAGGAGTCGAACCCCTGGCCTTTCGGCCAC

Ser_8967from *Chamaesiphon minutus*, anticodon: GCT, Seq. ID: C131004547

Seq.: GACGGTATAGTTTAATTGGTGAAAATACAATTCTGCTAAGGTTGCGATGGTGGTTCGAGTCCGTCTACTGTCACCA

tREX probe:aaaaaaaaaaaaTGGTGACAGTAGACGGACTCGAACCACCATC

Ser_2915from *Nitratireductor pacificus*, anticodon: TGA, Seq. ID: W131045483

Seq.:
GTCCGCATGGCGGAATTGGTAGACGCAGCAGTCTTGAAAACTGAAGCCTGTCCTGGGCATGCCGGTTCGAATCCGGCTGCGGACACCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGTGTCCGCAGCCGGATTCGAACCGGCATGCCCAGGACAGGCT

Ser_2957from *Granulicella mallensis*, anticodon: GCT, Seq. ID: C121006521

Seq.: GAGCCAGTAGCTCAGTTGGATAGAGCATCGGACTGCTAATCTGGGGGTCGGCGGTTCGAATCCGCCCTGGCATTCCA

tREX probe:aaaaaaaaaaaaaTGGAATGCCAGGGCGGATTCGAACCGCCGACC

Ser_5752from *Sulfuricurvum kujiense*, anticodon: GCT, Seq. ID: C11120697

Seq.:
GGACGAATGGGTGAGCGGCTGAAACCACCTCCCTGCTAAGGAGACGTACTGGCAACGGTACCGAGAGTTCAAATCTCTCTTCGTCCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGGACGAAGAGAGATTTGAACTCTCGGTACCGTTGCCAGTACG

Ser_3103from *Coprobacillus sp. 29_1*, anticodon: TGA, Seq. ID: W11119494

Seq.: GCTCAGTTGTGATAATTGGTAGTCGTGCTGTCTTGAAAACAGTTGGTCTGAAAGGACTTGGGGGTTCGAGTCCCTCACTGAGCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGCTCAGTGAGGGACTCGAACCCCCAAGTCCTTTCAGACC

Ser_8239from *Methanosarcina acetivorans*, anticodon: CGA, Seq. ID: At1411

Seq.: GCCCAGGTCGTCTAGTGGTATGGCGACGGTCTCGAAAACCGTTCCCTTTGGGATCGCAGGTTCAAATCCTGCCCTGGGCGCCA

tREX probe:aaaaaaaaaaaaaaaaaTGGCGCCCAGGGCAGGATTTGAACCTGCGATCCCAAAGGG

Ala_2698from *Tropheryma whipplei*, anticodon: CGC, Seq. ID: W131249897

Seq.: GGGCGATTGGCGCAGTGGTAGCGCGCTTCCATCGCACGGAAGAGGTCACTGGTTCGAGTCCAGTATCGCCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGGCGATACTGGACTCGAACCAGTGACC

Ala_3524from *Streptomyces sp. FXJ7.023*, anticodon: CGC, Seq. ID: W131109963

Seq.: ACGCTTGTAGCTCAGTGGATAGAGCAGCCCTTTCGCGAAGGGCAGGTCGCGGGTTCGAGTCCCGCCAGGCCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGGCCTGGCGGGACTCGAACCCGCGACC

Ala_2623from *Streptomyces sp. 142MFCol3.1*, anticodon: GGC, Seq. ID: W131234499

Seq.: CCCGGCGTCGAGGAGGGGCGGACTCGCCGCCTTGGCGAGGCGGATACGCCGGTTCCAATCCGGTCGTCGGGGCCA

tREX probe:aaaaaaaaaaaaTGGCCCCGACGACCGGATTGGAACCGGCGTA

Ala_3269from *Streptomyces purpureus*, anticodon: CGC, Seq. ID: W131163902

Seq.: CCGGATGTAGCGCAGAGGCAGGCGCACCGGTCTCGCAGGCCGGGCACGCCGGTTCGAATCCGGCCGTCGGGCCCA

tREX probe:aaaaaaaaaaaaTGGGCCCGACGGCCGGATTCGAACCGGCGTG

Ala_79from *Kitasatospora setae*, anticodon: GGC, Seq. ID: C11111760

Seq.: CCCGGGCGTGGGCGCAGAGGCCGCCGCCGCCGTTTTGGCAAAACGGAGCACGCCGGTTCGAATCCGGCCCCCGGGACCA

tREX probe:aaaaaaaaaaaaaTGGTCCCGGGGGCCGGATTCGAACCGGCGTGC

Ala_3046from *Salinispora arenicola*, anticodon: AGC, Seq. ID: W131177388

Seq.: GGGCCCGTAGCGCAGAGGTAGTCGCGCCGCCATAGCGTGGCGGAGGATGCCGGTTCGAGTCCGGTCGGGTCCCCCA

tREX probe:aaaaaaaaaaaaaTGGGGGACCCGACCGGACTCGAACCGGCATCC

Ala_3581from *Streptomyces sp. 351MFTsu5.1*, anticodon: GGC, Seq. ID: W131185494

Seq.: GGGTCGGTAGCGCAGAGGCAGGCGCGCCTTCATGGCATGAAGGGAACGTCGGTTCGAATCCGACCCGGCCCACCA

tREX probe:aaaaaaaaaaaaTGGTGGGCCGGGTCGGATTCGAACCGACGTT

Ala_1718from *Streptosporangium roseum*, anticodon: GGC, Seq. ID: C10112406

Seq.: GGGTCTGTAGCGCAGTCGGCCGTCGCGCCTCTTTGGCAAAGAGGAGGACGCTGGTTCGAATCCAGCCGGGCCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGGCCCGGCTGGATTCGAACCAGCGTCC

Ala_4032from *Actinoplanes sp. SE50/110*, anticodon: CGC, Seq. ID: C121007290

Seq.: GGGCCGGTAGCGCAGCTGGTCGACGCACACCGTTCGCATCGGTGGGGACGCCGGTTCGAATCCGGCTCGGCTCCCCA

tREX probe:aaaaaaaaaaaaaTGGGGAGCCGAGCCGGATTCGAACCGGCGTCC

Ala_2727from *Bacillus cereus*, anticodon: GGC, Seq. ID: C09102186

Seq.: GATCCCGTAGCTCAGCAGGGAGAGCGCCACCTTGGCAGGGTGGAGGTCGTGAGTTCGAGCCTCTCCGGGGTCACCA

tREX probe:aaaaaaaaaaaaaTGGTGACCCCGGAGAGGCTCGAACTCACGACC

Ala_3530from *Thermaerobacter marianensis*, anticodon: GGC, Seq. ID: C11122791

Seq.: GCGCCGGTAGCTCAGTTGGAAGAGCGGTGTTGTGGCGAAACACCAGGTCGAGGGTTCGAGTCCCTCCCGGCGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCGCCGGGAGGGACTCGAACCCTCGACCT

Ala_2780from *Gemella haemolysans*, anticodon: TGC, Seq. ID: W09119662

Seq.: GCCGGCCTAGCTCAGTTGGTAGAGCAACTGACTTGCAATCAGTAGGTCGGGGGTTCAAGTCCTCTGGCCGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCGGCCAGAGGACTTGAACCCCCGACC

Ala_3354from *Deinococcus sp. 2009*, anticodon: GGC, Seq. ID: W131213919

Seq.: GGTCCGGTAGCTCAGTGGAAGAGCATCTCACTGGCAGTGAGGGGGCCAGCGGTTCAATTCCGCTCCGGACGACCA

tREX probe:aaaaaaaaaaaaaTGGTCGTCCGGAGCGGAATTGAACCGCTGGCC

Ala_2762from *Acinetobacter baumannii*, anticodon: GGC, Seq. ID: W10106539

Seq.: GCAGCGGTAGTTCAGTTGGTTAGAATACCGGCCTGGCACGCCGGGGGTCGCGGGTTCGAGCCCCGTCCGCTGCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGCAGCGGACGGGGCTCGAACCCGCGACC

Ala_3674from *Sinorhizobium fredii*, anticodon: GGC, Seq. ID: W121066281

Seq.: GCGGGTGTAGCTCAGTCGGTTAGAGTGCCGGCGTGGCCCGCAGGAGGTCGCGGGTTCGAGCCCCGTCACTCGCGCCA

tREX probe:aaaaaaaaaaaaaTGGCGCGAGTGACGGGGCTCGAACCCGCGACC

Leu_7928from *Rhodobacter sphaeroides*, anticodon: TAG, Seq. ID: C019351

Seq.: CCTTTCGTAGCTCAGCTGGATAGAGCACCGGTCTTAGAAACCGGAGGTCGCAGGTTCGAGTCCTGCCGGGAGATCCA

tREX probe:aaaaaaaaaaaaaTGGATCTCCCGGCAGGACTCGAACCTGCGACC

Leu_12495from *Cenarchaeum symbiosum*, anticodon: GAG, Seq. ID: At0085

Seq.:
GCGGGTGTAGCCCAGCCTGGTCAAAGGCGCTAGCTTGAGGGGCTAGTCTCTTAGGAGTTCGTGGGTTCAAATCCCATCGCCCGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCGGGCGATGGGATTTGAACCCACGAACTCCTAAGAG

Leu_5765from *Methanosphaera stadtmanae*, anticodon: TAA, Seq. ID: At1577

Seq.: GCAGGGGTGCCCGAGCTGGCCAAAGGGGGTGGACTTAAGATCCTCTGGCGTAGGCCTACGTGGGTTCAAATCCCATCTCCTGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCAGGAGATGGGATTTGAACCCACGTAGGCCTACGCC

Leu_13580from *Methanothermobacter thermautotrophicus*, anticodon: TAG, Seq. ID: At1678

Seq.: GCGGGGGTGCCCGAGCTGGCCAAAGGGGACAGGCTTAGGACCTGTTGGCGTAGGCCTACCAGGGTTCGAATCCCTGCCCCCGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCGGGGGCAGGGATTCGAACCCTGGTAGGCCTACGCC

Leu_11584from *Pyrobaculum calidifontis*, anticodon: CAA, Seq. ID: At0315

Seq.:
GCGGGGGTGCCCGAGCCAGGCCAAAGGGGCAGGGCTCAAGACCCTGTGGCGTAGGCCTACGTGGGTTCAAATCCCACCCCCCGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCGGGGGGTGGGATTTGAACCCACGTAGGCCTACGCC

Leu_8101from *Sulfolobus acidocaldarius*, anticodon: CAA, Seq. ID: At0434

Seq.: GCGGGGGTGCCCGAGCTGGTCAAAGGGGGCGGACTCAAGATCCGCTGGCGAAGGCCTACGCGGGTTCAAATCCCGTCCCCCGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCGGGGGACGGGATTTGAACCCGCGTAGGCCTTCGCC

Chapter V – Appendix

Leu_8553from *Methanocaldococcus jannaschii*, anticodon: TAG, Seq. ID: At0932

Seq.:
GCAGGGGTCGCCAAGCCTGGCCAAAGGCGCTGGGCCTAGGACCCAGTCCCGTAGGGGTTCCAGGGTTCAAATCCCTGCCCCTGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCAGGGGCAGGGATTTGAACCCTGGAACCCCTACGGG

Leu_3196from *Methanococcus aeolicus*, anticodon: TAG, Seq. ID: At1021

Seq.:
GCAGGGGTTGTCGAGCCTGGCCAAAGACGCAGGACTTAGAATCCTGTCCTATAGTGGTTCCAGGGTTCAAATCCCTGCCCCTGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCAGGGGCAGGGATTTGAACCCTGGAACCACTATAGG

Leu_6706from *Methanoculleus marisnigri*, anticodon: CAG, Seq. ID: At1256

Seq.:
GCGAGAGTTGCCAAGCCAGGTCAAAGGCGCCAGGTTCAGGGCCTGGTCTCGTAGGAGTTCGCCGGTTCGAATCCGGCCTCTCGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCGAGAGGCCGGATTCGAACCGGCGAACTCCTACGAG

Leu_5696from *Methanosarcina acetivorans*, anticodon: CAG, Seq. ID: At1393

Seq.:
GCGAGAGTTGCCCAGCCAGGTCAAAGGCGCCAGGTTCAGAGCCTGGTCTTGTAGGAGTTCGTGCGTTCGAATCGCACCTCTCGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCGAGAGGTGCGATTCGAACGCACGAACTCCTACAAG

Leu_12008from *Methanosarcina barkeri*, anticodon: CAG, Seq. ID: At1493

Seq.:
GCGAGAGTTGCCCAGCCAGGTCAAAGGCGCCAGGTTCAGAGCCTGGTTTCGTAGGAATTCGTGCGTTCGAATCGCACCTCTCGCACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTGCGAGAGGTGCGATTCGAACGCACGAATTCCTACGAA

Leu_10651from *Holospora undulata*, anticodon: GAG, Seq. ID: W131181822

Seq.: GCGGTCATGGCGAAGCTGGTAGATGCGCAACGTTGAGGTCGTTGTGGGGGAAACCCCAGGGAAGTTCGAATCTTCTTGACCGCACCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGTGCGGTCAAGAAGATTCGAACTTCCCTGGGGTTTCCCCC

Leu_220from *Streptomyces avermitilis*, anticodon: CAA, Seq. ID: C020580

Seq.: GCCCCGGTAGTCCAGCGGTAGAGACACGGTGCTCAAACCACCGACAGCGTCGGTTCGAATCCGACTCGGGGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCCCCGAGTCGGATTCGAACCGACGCTG

Leu_12217from *Nocardiopsis gilva*, anticodon: TAG, Seq. ID: W131055138

Seq.: GGGGCCGTAGCCCAACAGGTAGAGGCACACGGTTTAGGTCCGTGCCAGCGCGGGTTCGAATCCCGCCGGCCCTACCA

tREX probe:aaaaaaaaaaaaaTGGTAGGGCCGGCGGGATTCGAACCCGCGCTG

Leu_7094from *Butyrivibrio sp. XBB1001*, anticodon: TAG, Seq. ID: W131233448

Seq.: GCACGAGTGGCGAAGCTGGTAGACGCACCTGACTTAGAATCAGACGCAGTAATGCATGTGGGTTCGAATCCCACCTCGTGTACCA

tREX probe:aaaaaaaaaaaaaaaaTGGTACACGAGGTGGGATTCGAACCCACATGCATTACTGCG

Leu_9241from *Candidatus Caldatribacteirum californiense*, anticodon: CAG, Seq. ID: W131110600

Seq.:
GCCGAAGTGGCGGAACTGGCAGACGCGCTGGAATCAGGCTCCAGTGAGCGTAAAGCTCATGCGGGTTCGACTCCCGCCTTCGGCGCCA

tREX probe:aaaaaaaaaaaaaaaaaaTGGCGCCGAAGGCGGGAGTCGAACCCGCATGAGCTTTACGCTC

Leu_13688from *Nocardioides sp. Iso805N*, anticodon: TAG, Seq. ID: W131155928

Seq.: GCGCCCGTAGCCCAACTGGCAGAGGCATACGGTTTAGGTCCGTACCAGCGTGAGTTCGAATCTCACCGGGCGCACCA

tREX probe:aaaaaaaaaaaaaTGGTGCGCCCGGTGAGATTCGAACTCACGCTG

Leu_8822from *Opitutus terrae*, anticodon: CAG, Seq. ID: C08012631

Seq.: CGGCGAGTGGTGGAGCGTATACACACGAGTCTCAGAAACTCGCGGGCGAAAGCCCATGCAGGTGCAAGTCCTGTCTCGCCGACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTCGGCGAGACAGGACTTGCACCTGCATGGGCTTTCGCCC

Leu_10255from *Pseudopropionibacterium propionicum*, anticodon: TAA, Seq. ID: C121002928

Seq.: GCCCCCGTAGCCCAACTGGTAGAGGCAGGCGGCTTAAACCCGCCCCAGTGCGGGTTCGAGTCCCGTCGGGGGTGCCA

tREX probe:aaaaaaaaaaaaaTGGCACCCCCGACGGGACTCGAACCCGCACTG

Leu_7797from *Proteobacteria bacterium JGI 0000113-P07*, anticodon: TAA, Seq. ID: W131237883

Seq.: GCGGGGGTCGCATAGTCTGGCCAATTGCGCCGGACTTAAGATCCGGTCCTTAGTGGTACGAGGGTTCGAATCCCTCCCCCCGCACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTGCGGGGGGAGGGATTCGAACCCTCGTACCACTAAGG

Leu_8506from *Thermobispora bispora*, anticodon: TAG, Seq. ID: C10113113

Seq.: GGGGTCGTAGCCCAACCGGCAGAGGCGACGGTTTTAGGTACCGTACAGCGCGAGTTCGAATCTCGCCGACCCCACCA

tREX probe:aaaaaaaaaaaaaTGGTGGGGTCGGCGAGATTCGAACTCGCGCTG

Leu_6101from *Thioalkalivibrio thiocyanodenitrificans*, anticodon: TAA, Seq. ID: W131163199

Seq.: GCTCCGGTGGCGAAGCTGGTAGACGCAGGGCACTTAAAATGCCCCGCCGGCAGGCATGCGGGTTCGACTCCCGCCCGGAGCACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTGCTCCGGGCGGGAGTCGAACCCGCATGCCTGCCGGC

Leu_12207from *Haloarcula marismortui*, anticodon: GAG, Seq. ID: At0775

Seq.:
GCGTGGGTAGCCAAGCCAGGCCAACGGCGCAGCGTTGAGGGCGCTGTCCCGTAGGGGTCCGCCGGTTCGAATCCGGTCCCACGCACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTGCGTGGGACCGGATTCGAACCGGCGGACCCCTACGGG

Leu_8812from *Haloquadratum walsbyi*, anticodon: GAG, Seq. ID: At0877

Seq.:
GCGTGGGTAGCCAAGCCAGGCCAACGGCGCAGCGTTGAGGGCGCTGTCCTGTAGAGGTCCGCCGGTTCAAATCCGGTCCCACGCACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTGCGTGGGACCGGATTTGAACCGGCGGACCTCTACAGG

Leu_6208from *Natronomonas pharaonis*, anticodon: GAG, Seq. ID: At1693

Seq.:
GCGTGGGTAGCCAAGCTAGGCCAACGGCGCAGCGTTGAGGGCGCTGTCCTGTAGAGGTCCGCCGGTTCAAATCCGGTCCCACGCACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTGCGTGGGACCGGATTTGAACCGGCGGACCTCTACAGG

Leu_7869from *Halobacterium salinarum*, anticodon: GAG, Seq. ID: At0829

Seq.:
GCGTGGGTAGCCGAGCTAGGTCAAAGGCGCAGCGTTGAGGGCGCTGTCCTGTAGAGGTTCGCCGGTTCGAATCCGGTCCCACGCACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTGCGTGGGACCGGATTCGAACCGGCGAACCTCTACAGG

Leu_6187from *Archaeoglobus fulgidus*, anticodon: GAG, Seq. ID: At0608

Seq.: GCGGGGGTTGCCGAGCGGACAAAGGCGCAGGATTGAGGGTCCTGTCCCGTAGGGGTTCGAGGGTTCGAATCCCTCCCCCCGCACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTGCGGGGGGAGGGATTCGAACCCTCGAACCCCTACGGG

Leu_9173from *Methanoregula boonei*, anticodon: GAG, Seq. ID: At0658

Seq.:
GCGAGGGTAGCCAAGCCAGGTCAAAGGCGCTAGGTTGAGGGCCTAGTCTCGTAGGAGTTCTTGGGTTCGAATCCCATCCCTCGCACCA

tREX probe:aaaaaaaaaaaaaaaaaTGGTGCGAGGGATGGGATTCGAACCCAAGAACTCCTACGAG

D Loop Alignment Table

Below, the alignment of the D loop to the canonical positions 14→21, which was manually generated, is indicated.

AAAGCA→AA--AGCA
GTGGCA→GT--GGCA
AAAGGCA→AAA-GGCA
AAAGGTA→AAA-GGTA
AACAGGC→AACAGG-C
AACGGAT→AAC-GGAT
AACGGGA→AAC-GGGA
AACGTAG→AAC-GTAG
AAGGATA→AA--GGAA
AAGGGGA→AAG-GGGA
AATAGAA→AAT-AGAA
AATGAGA→AAT-GAGA
AATGCTA→AAT-GCTA
AATGGCA→AAT-GGCA
AATGGGT→AAT-GGGT
AATGTAG→AAT-GTAG
AATTGGA→AATTGG-A
ACCCGGA→ACCCGG-A
ACCGGTC→ACC-GGTC
ACTGGAA→ACT-GGAA
ACTGGTT→ACT-GGTT
AGACGTA→AGA-CGTA
AGAGGCA→AGA-GGCA
AGAGGGT→AGA-GGGT
AGAGTTT→AGA-GTTT
AGCGACC→AGC-GACC
AGCGGCG→AGC-GGCG
AGCGGTA→AGC-GGTA
AGCGTTA→AGC-GTTA
AGCTGTA→AGC-TGTA
AGGGGCA→AGG-GGCA
AGGGGTC→AGG-GGTC
AGTAATT→AGT-AATT
AGTGACA→AGT-GACA
AGTGATT→AGT-GATT
AGTGGAT→AGT-GGAT
AGTGGGA→AGT-GGGA
AGTGTTA→AGT-GTTA
ANCGGTA→ANC-GGTA
ATCGGAA→ATC-GGAA
ATCGGCG→ATC-GGCG
ATCGGTT→ATC-GGTT
ATGGTGC→AT--GGTC
ATTGGGA→ATT-GGGA
ATTGGTT→ATT-GGTT
CACGGCG→CAC-GGCG
CGCGGCC→CGC-GGCC
CTGGGAG→CTG-GGAG
GAAGGAC→GAA-GGAC
GACGGCC→GAC-GGCC
GAGGGAG→GAG-GGAG
GATGGCA→GAT-GGCA
GATGGTA→GAT-GGTA
GCTGGCT→GCT-GGCT
GGCGGTG→GGC-GGTG

AGCTGG→AGCTGG--
AAAAGCA→AAA-AGCA
AAAGGCC→AAA-GGCC
AAAGGTG→AAA-GGTG
AACAGTA→AAC-AGTA
AACGGCA→AAC-GGCA
AACGGTA→AAC-GGTA
AACTCTA→AAC-TCTA
AAGGGAA→AAG-GGAA
AAGGGTA→AAG-GGTA
AATAGCA→AAT-AGCA
AATGATA→AAT-GATA
AATGGAA→AAT-GGAA
AATGGCC→AAT-GGCC
AATGGTA→AAT-GGTA
AATGTTA→AAT-GTTA
AATTGGC→AATTGG-C
ACCGATA→ACC-GATA
ACCGGTG→ACC-GGTG
ACTGGCA→ACT-GGCA
AGAAGAT→AGA-AGAT
AGAGAAA→AGA-GAAA
AGAGGCC→AGA-GGCC
AGAGGTA→AGA-GGTA
AGCAGCA→AGC-AGCA
AGCGGAA→AGC-GGAA
AGCGGCT→AGC-GGCT
AGCGGTC→AGC-GGTC
AGCTGAA→AGCTGA-A
AGGAGCA→AGG-AGCA
AGGGGCC→AGG-GGCC
AGGGGTG→AGG-GGTG
AGTAGCA→AGT-AGCA
AGTGACC→AGT-GACC
AGTGCAA→AGT-GCAA
AGTGGCA→AGT-GGCA
AGTGGTA→AGT-GGTA
AGTTAGG→AGTTAG-G
ANGCGGT→ANGCGG-T
ATCGGAG→ATC-GGAG
ATCGGCT→ATC-GGCT
ATCTGGA→ATCTGG-A
ATTGGAA→ATT-GGAA
ATTGGGG→ATT-GGGG
CAAAGTA→CAA-AGTA
CATGGCG→CAT-GGCG
CGCGGTA→CGC-GGTA
CTGGTGC→CT--GGTC
GAAGGAT→GAA-GGAT
GACGGTA→GAC-GGTA
GAGGGCA→GAG-GGCA
GATGGCC→GAT-GGCC
GATGGTG→GAT-GGTG
GCTGGGA→GCT-GGGA
GGCGGTT→GGC-GGTT

AGGGTA→AG--GGTA
AAAGCAG→AAA-GCAG
AAAGGCG→AAA-GGCG
AAAGGTT→AAA-GGTT
AACGATA→AAC-GATA
AACGGCC→AAC-GGCC
AACGGTC→AAC-GGTC
AACTGCA→AAC-TGCA
AAGGGAG→AAG-GGAG
AAGGGTG→AAG-GGTG
AATAGTA→AAT-AGTA
AATGATG→AAT-GATG
AATGGAC→AAT-GGAC
AATGGCG→AAT-GGCG
AATGGTC→AAT-GGTC
AATTATA→AAT-TATA
AATTGGT→AATTGG-T
ACCGGCA→ACC-GGCA
ACGGCAG→AC--GGCG
ACTGGCT→ACT-GGCT
AGAAGCA→AGA-AGCA
AGAGATG→AGA-GATG
AGAGGCG→AGA-GGCG
AGAGGTC→AGA-GGTC
AGCAGCG→AGC-AGCG
AGCGGAT→AGC-GGAT
AGCGGGA→AGC-GGGA
AGCGGTG→AGC-GGTG
AGCTGGA→AGCTGG-A
AGGAGTA→AGG-AGTA
AGGGGCT→AGG-GGCT
AGGGGTT→AGG-GGTT
AGTAGGT→AGTAGG-T
AGTGATA→AGT-GATA
AGTGCTA→AGT-GCTA
AGTGGCC→AGT-GGCC
AGTGGTC→AGT-GGTC
AGTTGGA→AGTTGG-A
ATAGGTA→ATA-GGTA
ATCGGAT→ATC-GGAT
ATCGGTA→ATC-GGTA
ATGGGTA→ATG-GGTA
ATTGGAT→ATT-GGAT
ATTGGTA→ATT-GGTA
CAAGGCG→CAA-GGCG
CATGGTG→CAT-GGTG
CGGGGTG→CGG-GGTG
CTGGTGG→CT--GGTG
GAAGGCG→GAA-GGCG
GACGGTC→GAC-GGTC
GATAGGA→GATAGG-A
GATGGCG→GAT-GGCG
GCAGGCT→GCA-GGCT
GCTGGTA→GCT-GGTA
GGGAGCA→GGG-AGCA

ATGGTG→AT--GGTG
AAAGGAA→AAA-GGAA
AAAGGCT→AAA-GGCT
AACAGCA→AAC-AGCA
AACGCAG→AAC-GCAG
AACGGCG→AAC-GGCG
AACGGTG→AAC-GGTG
AACTTTA→AAC-TTTA
AAGGGAT→AAG-GGAT
AAGGGTT→AAG-GGTT
AATAGTG→AAT-AGTG
AATGCAG→AAT-GCAG
AATGGAG→AAT-GGAG
AATGGCT→AAT-GGCT
AATGGTG→AAT-GGTG
AATTCTA→AAT-TCTA
AATTTTA→AAT-TTTA
ACCGGCC→ACC-GGCC
ACGGGTA→ACG-GGTA
ACTGGGA→ACT-GGGA
AGAAGTA→AGA-AGTA
AGAGGAA→AGA-GGAA
AGAGGCT→AGA-GGCT
AGAGGTG→AGA-GGTG
AGCAGTA→AGC-AGTA
AGCGGCA→AGC-GGCA
AGCGGGG→AGC-GGGG
AGCGGTT→AGC-GGTT
AGCTGGG→AGCTGG-G
AGGGATA→AG--GGAA
AGGGGGA→AGG-GGGA
AGGGTTA→AG--GGTA
AGTAGTA→AGT-AGTA
AGTGGTT→AGTTGG-T
ATAGGTG→ATA-GGTG
ATCGGCA→ATC-GGCA
ATCGGTC→ATC-GGTC
ATGGGTT→ATG-GGTT
ATTGGCA→ATT-GGCA
ATTGGTC→ATT-GGTC
CAAGGTG→CAA-GGTG
CCAAGTC→CCA-AGTC
CGTGGTA→CGT-GGTA
CTGGTGT→CT--GGTT
GAAGGTC→GAA-GGTC
GACGGTG→GAC-GGTG
GATGCAG→GAT-GCAG
GATGGCT→GAT-GGCT
GCCGGAT→GCC-GGAT
GGAGGTT→GGA-GGTT
GGTGGAC→GGT-GGAC

CTGGTA→CT--GGTA
AAAGGAT→AAA-GGAT
AAAGGGA→AAA-GGGA
AACAGGA→AACAGG-A
AACGGAA→AAC-GGAA
AACGGCT→AAC-GGCT
AACGGTT→AAC-GGTT
AAGAGTA→AAG-AGTA
AAGGGCA→AAG-GGCA
AAGGTCA→AA--GGTA
AATGAAG→AAT-GAAG
AATGCAT→AAT-GCAT
AATGGAT→AAT-GGAT
AATGGGA→AAT-GGGA
AATGGTT→AAT-GGTT
AATTGAG→AATTGA-G
ACAGGTA→ACA-GGTA
ACCGGTA→ACC-GGTA
ACGGGTG→ACG-GGTG
ACTGGTA→ACT-GGTA
AGAAGTC→AGA-AGTC
AGAGGAT→AGA-GGAT
AGAGGGA→AGA-GGGA
AGAGGTT→AGA-GGTT
AGCAGTT→AGC-AGTT
AGCGGCC→AGC-GGCC
AGCGGGT→AGC-GGGT
AGCGTAT→AGC-GTAT
AGCTGGT→AGCTGG-T
AGGGGAA→AGG-GGAA
AGGGGTA→AGG-GGTA
AGNGGTA→AGN-GGTA
AGTCCGG→AGT-CCGG
AGTGATG→AGT-GATG
AGTGGAG→AGT-GGAG
AGTGGCT→AGT-GGCT
AGTGGTT→AGT-GGTT
AGTTGTA→AGTTGT-A
ATAGGTT→ATA-GGTT
ATCGGCC→ATC-GGCC
ATCGGTG→ATC-GGTG
ATGGTAG→AT--GGTG
ATTGGCT→ATT-GGCT
ATTGGTG→ATT-GGTG
CAATGCA→CAA-TGCA
CCGGTGC→CC--GGTC
CTGGATG→CT--GGAG
CTTGGTA→CTT-GGTA
GACGGCA→GAC-GGCA
GACGGTT→GAC-GGTT
GATGGAT→GAT-GGAT
GATGGGA→GAT-GGGA
GCCGGTA→GCC-GGTA
GGCGGTA→GGC-GGTA
GGTGGAT→GGT-GGAT

GGTGGTA→GGT-GGTA
GTCGGTT→GTC-GGTT
GTTGGAT→GTT-GGAT
TCAGCTG→TCA-GCTG
TGCGGCA→TGC-GGCA
TGTGGTT→TGT-GGTT
AAAAGCCA→AAAAGCCA
AAACGGCT→AAACGGCT
AAAGGACA→AAA-GGAA
AAAGGCAA→AAA-GGCA
AAAGGCTA→AAA-GGCA
AAAGGTAG→AAA-GGTG
AAATGGAA→AAATGGAA
AAATGGCT→AAATGGCT
AAATGGTG→AAATGGTG
AACAGGCA→AACAGGCA
AACCAGCA→AACCAGCA
AACCGGGA→AACCGGGA
AACGATTA→AAC-GATA
AACGGAAT→AAC-GGAT
AACGGATC→AAC-GGAC
AACGGCCA→AAC-GGCA
AACGGCTT→AAC-GGCT
AACGGGCT→AACGGGCT
AACGGTCA→AAC-GGTA
AACGGTTT→AAC-GGTT
AACTGCCA→AACTGCCA
AACTGGCC→AACTGGCC
AACTGGTA→AACTGGTA
AACTTGTA→AACTTGTA
AAGCGGAA→AAGCGGAA
AAGGGACT→AAG-GGAT
AAGGGCTA→AAG-GGCA
AAGGGTAG→AAG-GGTG
AAGTGGCG→AAGTGGCG
AATAGAAA→AATAGAAA
AATAGGCA→AATAGGCA
AATCAGCA→AATCAGCA
AATCGGTA→AATCGGTA
AATGATTA→AAT-GATA
AATGGACA→AAT-GGAA
AATGGATG→AAT-GGAG
AATGGCCA→AAT-GGCA
AATGGCTG→AAT-GGCG
AATGGGCA→AATGGGCA
AATGGTAT→AAT-GGTT
AATGGTTC→AAT-GGTC
AATGTGTA→AATGTGTA
AATTCTTA→AATTCTTA
AATTGGAT→AATTGGAT
AATTGGTA→AATTGGTA
AATTTCTA→AATTTCTA
ACAGGTTA→ACA-GGTA
ACCCGGTG→ACCCGGTG
ACCGGTCA→ACC-GGTA
ACCTGGCT→ACCTGGCT
ACGCGGTA→ACGCGGTA
ACTAGGTA→ACTAGGTA
ACTGGCGT→ACT-GGCT
ACTGGTTT→ACT-GGTT
AGAAGGAA→AGAAGGAA
AGACGGTA→AGACGGTA
AGAGGAAA→AGA-GGAA
AGAGGCAA→AGA-GGCA
AGAGGCCT→AGA-GGCT
AGAGGCTT→AGA-GGCT
AGAGGTCG→AGA-GGTG
AGAGGTTG→AGA-GGTG
AGATGGAA→AGATGGAA
AGATGGTA→AGATGGTA
AGCAGCGA→AGCAGCGA
AGCAGGCG→AGCAGGCG

GGTGGTT→GGT-GGTT
GTGGCAA→GT--GGCA
GTTGGTA→GTT-GGTA
TCCGGGA→TCC-GGGA
TGCGGTG→TGC-GGTG
TTAGAAA→TTA-GAAA
AAAAGGGA→AAAAGGGA
AAACGGGA→AAACGGGA
AAAGGAGA→AAA-GGAA
AAAGGCAG→AAA-GGCG
AAAGGGAC→AAAGGGAC
AAAGGTCA→AAA-GGTA
AAATGGAC→AAATGGAC
AAATGGGA→AAATGGGA
AAATGGTT→AAATGGTT
AACAGGGA→AACAGGGA
AACCGGAA→AACCGGAA
AACCGGTA→AACCGGTA
AACGATTT→AAC-GATT
AACGGACA→AAC-GGAA
AACGGATT→AAC-GGAT
AACGGCCG→AAC-GGCG
AACGGGAA→AACGGGAA
AACGGGTA→AACGGGTA
AACGGTGA→AAC-GGTA
AACTCTAA→AACTCTAA
AACTGGAA→AACTGGAA
AACTGGCG→AACTGGCG
AACTGGTC→AACTGGTC
AACTTTCA→AACTTTCA
AAGCGGCA→AAGCGGCA
AAGGGATA→AAG-GGAA
AAGGGCTG→AAG-GGCG
AAGGGTTA→AAG-GGTA
AAGTGGGA→AAGTGGGA
AATAGCAG→AATAGCAG
AATAGGGA→AATAGGGA
AATCGCTA→AATCGCTA
AATCGGTG→AATCGGTG
AATGGAAA→AAT-GGAA
AATGGAGA→AAT-GGAA
AATGGATT→AAT-GGAT
AATGGCGA→AAT-GGCA
AATGGCTT→AAT-GGCT
AATGGGGA→AATGGGGA
AATGGTCA→AAT-GGTA
AATGGTTG→AAT-GGTG
AATTACAA→AATTACAA
AATTGACA→AATTGACA
AATTGGCA→AATTGGCA
AATTGGTC→AATTGGTC
AATTTGTA→AATTTGTA
ACCAGGCA→ACCAGGCA
ACCGGCGA→ACC-GGCA
ACCGGTGA→ACC-GGTA
ACCTGGGA→ACCTGGGA
ACGGGATA→ACG-GGAA
ACTCGGTA→ACTCGGTA
ACTGGCTA→ACT-GGCA
ACTTGGCA→ACTTGGCA
AGAAGGTA→AGAAGGTA
AGACGGTC→AGACGGTC
AGAGGACA→AGA-GGAA
AGAGGCAG→AGA-GGCG
AGAGGCGG→AGA-GGCG
AGAGGGTA→AGA-GGGA
AGAGGTGA→AGA-GGTA
AGAGGTTT→AGA-GGTT
AGATGGAG→AGATGGAG
AGATGGTC→AGATGGTC
AGCAGCTA→AGCAGCTA
AGCAGGGA→AGCAGGGA

GGTTGTA→GGT-TGTA
GTGGGAC→GT--GGGC
TAAGGCA→TAA-GGCA
TCTGGGA→TCT-GGGA
TGTAGAT→TGT-AGAT
TTAGGTA→TTA-GGTA
AAAAGGTA→AAAAGGTA
AAACGGTA→AAACGGTA
AAAGGATA→AAA-GGAA
AAAGGCAT→AAA-GGCT
AAAGGGAG→AAAGGGAG
AAAGGTGA→AAA-GGTA
AAATGGAT→AAATGGAT
AAATGGGG→AAATGGGG
AACAGCAG→AACAGCAG
AACAGGTA→AACAGGTA
AACCGGAG→AACCGGAG
AACGAATA→AAC-GAAA
AACGCTCA→AAC-GCTA
AACGGACG→AAC-GGAG
AACGGCAA→AAC-GGCA
AACGGCGA→AAC-GGCA
AACGGGAC→AACGGGAC
AACGGTAA→AAC-GGTA
AACGGTTA→AAC-GGTA
AACTCTCA→AACTCTCA
AACTGGAC→AACTGGAC
AACTGGCT→AACTGGCT
AACTGGTT→AACTGGTT
AACTTTTA→AACTTTTA
AAGCGGTA→AAGCGGTA
AAGGGATG→AAG-GGAG
AAGGGCTT→AAG-GGCT
AAGTGGAA→AAGTGGAA
AAGTGGTA→AAGTGGTA
AATAGCCA→AATAGCCA
AATAGGTA→AATAGGTA
AATCGGAA→AATCGGAA
AATGAATA→AAT-GAAA
AATGGAAC→AAT-GGAC
AATGGAGC→AAT-GGAC
AATGGCAA→AAT-GGCA
AATGGCGG→AAT-GGCG
AATGGGAA→AATGGGAA
AATGGGTA→AAT-GGGA
AATGGTGA→AAT-GGTA
AATGGTTT→AAT-GGTT
AATTACCA→AATTACCA
AATTGATA→AATTGATA
AATTGGCG→AATTGGCG
AATTGGTG→AATTGGTG
AATTTTTA→AATTTTTA
ACCAGGTA→ACCAGGTA
ACCGGCTA→ACC-GGCA
ACCGGTGT→ACC-GGTT
ACCTGGTA→ACCTGGTA
ACGGGTTA→ACG-GGTA
ACTGGAGA→ACT-GGAA
ACTGGGGT→ACTGGGGT
ACTTGGGA→ACTTGGGA
AGACGGAA→AGACGGAA
AGACGTTT→AGACGTTT
AGAGGAGA→AGA-GGAA
AGAGGCCA→AGA-GGCA
AGAGGCTA→AGA-GGCA
AGAGGTAA→AGA-GGTA
AGAGGTGG→AGA-GGTG
AGAGTGTA→AGAGTGTA
AGATGGAT→AGATGGAT
AGATGGTT→AGATGGTT
AGCAGCTC→AGCAGCTC
AGCAGGTA→AGCAGGTA

GTAGGTT→GTA-GGTT
GTGGTAG→GT--GGTG
TACGGCA→TAC-GGCA
TCTGGTA→TCT-GGTA
TGTGGTA→TGT-GGTA
TTTGGCT→TTT-GGCT
AAACGGAA→AAACGGAA
AAACGTTA→AAACGTTA
AAAGGATC→AAA-GGAC
AAAGGCCA→AAA-GGCA
AAAGGGAT→AAAGGGAT
AAAGGTTA→AAA-GGTA
AAATGGCA→AAATGGCA
AAATGGTA→AAATGGTA
AACAGCAT→AACAGCAT
AACAGGTG→AACAGGTG
AACCGGCA→AACCGGCA
AACGAGCA→AAC-GAGA
AACGGAAA→AAC-GGAA
AACGGAGA→AAC-GGAA
AACGGCAG→AAC-GGCG
AACGGCTA→AAC-GGCA
AACGGGAG→AACGGGAG
AACGGTAG→AAC-GGTG
AACGGTTC→AAC-GGTC
AACTCTTA→AACTCTTA
AACTGGAT→AACTGGAT
AACTGGGA→AACTGGGA
AACTGTCA→AACTGTCA
AAGAGCTA→AAGAGCTA
AAGGGAAA→AAG-GGAA
AAGGGCAG→AAG-GGCG
AAGGGGTA→AAGGGGTA
AAGTGGAT→AAGTGGAT
AAGTGGTT→AAGTGGTT
AATAGCTA→AATAGCTA
AATAGTCA→AATAGTCA
AATCGGCA→AATCGGCA
AATGACCA→AAT-GACA
AATGGAAG→AAT-GGAG
AATGGATA→AAT-GGAA
AATGGCAG→AAT-GGCG
AATGGCTA→AAT-GGCA
AATGGGAG→AATGGGAG
AATGGTAA→AAT-GGTA
AATGGTGT→AAT-GGTT
AATGTCAG→AAT-GTCG
AATTAGCA→AATTAGCA
AATTGCAG→AATTGCAG
AATTGGCT→AATTGGCT
AATTGGTT→AATTGGTT
ACAAGGTA→ACAAGGTA
ACCCGGCA→ACCCGGCA
ACCGGGAA→ACCGGGAA
ACCGGTTA→ACC-GGTA
ACCTGGTT→ACCTGGTT
ACGTGGAA→ACGTGGAA
ACTGGATA→ACT-GGAA
ACTGGTAA→ACT-GGTA
ACTTGGTA→ACTTGGTA
AGACGGCA→AGACGGCA
AGAGAGAT→AGA-GAGT
AGAGGATA→AGA-GGAA
AGAGGCCC→AGA-GGCC
AGAGGCTC→AGA-GGCC
AGAGGTAG→AGA-GGTG
AGAGGTTA→AGA-GGTA
AGATAGCA→AGATAGCA
AGATGGCA→AGATGGCA
AGCAGCAA→AGCAGCAA
AGCAGGAA→AGCAGGAA
AGCAGGTG→AGCAGGTG

GTCGGTA→GTC-GGTA
GTTAGAG→GTT-AGAG
TACGGTA→TAC-GGTA
TGCGGAG→TGC-GGAG
TGTGGTG→TGT-GGTG
AAAACATA→AAAACATA
AAACGGCA→AAACGGCA
AAAGGAAA→AAA-GGAA
AAAGGATT→AAA-GGAT
AAAGGCGA→AAA-GGCA
AAAGGTAA→AAA-GGTA
AAATCTCA→AAATCTCA
AAATGGCG→AAATGGCG
AAATGGTC→AAATGGTC
AACAGGAA→AACAGGAA
AACAGTAG→AACAGTAG
AACCGGCG→AACCGGCG
AACGATGA→AAC-GATA
AACGGAAG→AAC-GGAG
AACGGATA→AAC-GGAA
AACGGCAT→AAC-GGCT
AACGGCTG→AAC-GGCG
AACGGGCA→AACGGGCA
AACGGTAT→AAC-GGTT
AACGGTTG→AAC-GGTG
AACTGATA→AACTGATA
AACTGGCA→AACTGGCA
AACTGGGT→AACTGGGT
AACTGTTA→AACTGTTA
AAGAGTTA→AAGAGTTA
AAGGGACA→AAG-GGAA
AAGGGCAT→AAG-GGCT
AAGGGTAA→AAG-GGTA
AAGTGGCA→AAGTGGCA
AANTGGTA→AANTGGTA
AATAGGAA→AATAGGAA
AATAGTTA→AATAGTTA
AATCGGGA→AATCGGGA
AATGACTA→AAT-GACA
AATGGAAT→AAT-GGAT
AATGGATC→AAT-GGAC
AATGGCAT→AAT-GGCT
AATGGCTC→AAT-GGCC
AATGGGAT→AATGGGAT
AATGGTAG→AAT-GGTG
AATGGTTA→AAT-GGTA
AATGTCTA→AAT-GTCA
AATTAGTA→AATTAGTA
AATTGGAA→AATTGGAA
AATTGGGA→AATTGGGA
AATTGTTA→AATTGTTA
ACAGGCAG→ACA-GGCG
ACCCGGTA→ACCCGGTA
ACCGGGTA→ACCGGGTA
ACCTGGCA→ACCTGGCA
ACGAGGTA→ACGAGGTA
ACTAGGCA→ACTAGGCA
ACTGGCAG→ACT-GGCG
ACTGGTGT→ACT-GGTT
AGAAGCTC→AGAAGCTC
AGACGGGA→AGACGGGA
AGAGCCTA→AGA-GCCA
AGAGGATC→AGA-GGAC
AGAGGCCG→AGA-GGCG
AGAGGCTG→AGA-GGCG
AGAGGTCA→AGA-GGTA
AGAGGTTC→AGA-GGTC
AGATCGGA→AGATGG-A
AGATGGGA→AGATGGGA
AGCAGCCA→AGCAGCCA
AGCAGGCA→AGCAGGCA
AGCAGTCA→AGCAGTCA

Chapter V – Appendix

AGCAGTGA→AGCAGTGA
AGCCAGGA→AGCCGG-A
AGCCGGAA→AGCCGGAA
AGCCGGCA→AGCCGGCA
AGCCGGCG→AGCCGGCG
AGCCGGGA→AGCCGGGA
AGCCGGTA→AGCCGGTA
AGCCTGGA→AGCCGG-A
AGCCTGGC→AGCCGG-C
AGCCTGGT→AGCCGG-T
AGCCTGTA→AGCCTGTA
AGCGAACA→AGC-GAAA
AGCGAGAA→AGC-GAGA
AGCGATTA→AGC-GATA
AGCGCACA→AGC-GCAA
AGCGGAAA→AGC-GGAA
AGCGGAAC→AGC-GGAC
AGCGGAAG→AGC-GGAG
AGCGGAAT→AGC-GGAT
AGCGGACA→AGC-GGAA
AGCGGAGA→AGC-GGAA
AGCGGAGT→AGC-GGAT
AGCGGATA→AGC-GGAA
AGCGGATC→AGC-GGAC
AGCGGATG→AGC-GGAG
AGCGGATT→AGC-GGAT
AGCGGCAA→AGC-GGCA
AGCGGCAC→AGC-GGCC
AGCGGCAG→AGC-GGCG
AGCGGCCA→AGC-GGCA
AGCGGCCC→AGC-GGCC
AGCGGCGA→AGC-GGCA
AGCGGCTA→AGC-GGCA
AGCGGCTC→AGC-GGCC
AGCGGCTG→AGC-GGCG
AGCGGCTT→AGC-GGCT
AGCGGGAG→AGCGGGAG
AGCGGGCA→AGCGGGCA
AGCGGGGA→AGCGGGGA
AGCGGGTA→AGC-GGGA
AGCGGGTG→AGC-GGGG
AGCGGTAA→AGC-GGTA
AGCGGTAC→AGC-GGTC
AGCGGTAG→AGC-GGTG
AGCGGTCA→AGC-GGTA
AGCGGTCG→AGC-GGTG
AGCGGTGA→AGC-GGTA
AGCGGTGG→AGC-GGTG
AGCGGTGT→AGC-GGTT
AGCGGTTA→AGC-GGTA
AGCGGTTC→AGC-GGTC
AGCGGTTG→AGC-GGTG
AGCGGTTT→AGC-GGTT
AGCGTAGA→AGC-GTAA
AGCGTGTA→AGCGTGTA
AGCTAGTA→AGCTAGTA
AGCTGAGA→AGCTGAGA
AGCTGATA→AGCTGATA
AGCTGGAA→AGCTGGAA
AGCTGGAG→AGCTGGAG
AGCTGGCA→AGCTGGCA
AGCTGGCT→AGCTGGCT
AGCTGGGA→AGCTGGGA
AGCTGGGC→AGCTGGGC
AGCTGGGG→AGCTGGGG
AGCTGGGT→AGCTGGGT
AGCTGGTA→AGCTGGTA
AGCTGGTG→AGCTGGTG
AGCTGGTN→AGCTGGTN
AGCTGGTT→AGCTGGTT
AGCTGTGA→AGCTGTGA
AGCTGTTA→AGCTGTTA
AGCTTGGA→AGCTGG-A
AGGAGATA→AGGAGATA
AGGAGGAA→AGGAGGAA
AGGAGGCA→AGGAGGCA
AGGAGGTA→AGGAGGTA
AGGCAGCG→AGGCAGCG
AGGCAGTA→AGGCAGTA
AGGCGGCA→AGGCGGCA
AGGCGGGA→AGGCGGGA
AGGCGGTA→AGGCGGTA
AGGCGGTG→AGGCGGTG
AGGCGGTT→AGGCGGTT
AGGGCTTA→AG--GGCA
AGGGGAAA→AGG-GGAA
AGGGGACA→AGG-GGAA
AGGGGAGA→AGG-GGAA
AGGGGATA→AGG-GGAA
AGGGGCAG→AGG-GGCG
AGGGGCCA→AGG-GGCA
AGGGGCGG→AGG-GGCG
AGGGGCTA→AGG-GGCA
AGGGGCTT→AGG-GGCT
AGGGGGAA→AGGGGGAA
AGGGGGAC→AGGGGGAC
AGGGGGAT→AGGGGGAT
AGGGGGCA→AGGGGGCA
AGGGGGTA→AGGGGGTA
AGGGGTAA→AGG-GGTA
AGGGGTAC→AGG-GGTC
AGGGGTCA→AGG-GGTA
AGGGGTGA→AGG-GGTA
AGGGGTTA→AGG-GGTA
AGGGGTTT→AGG-GGTT
AGGGTTTA→AG--GGTA
AGGNGGTA→AGGNGGTA
AGGTAGTA→AGGTAGTA
AGGTGGAA→AGGTGGAA
AGGTGGCA→AGGTGGCA
AGGTGGCG→AGGTGGCG
AGGTGGCT→AGGTGGCT
AGGTGGGA→AGGTGGGA
AGGTGGTA→AGGTGGTA
AGGTGGTC→AGGTGGTC
AGGTGGTG→AGGTGGTG
AGGTGGTT→AGGTGGTT
AGGTGTGA→AGGTGTGA
AGKTGGTA→AGKTGGTA
AGNTGGGA→AGNTGGGA
AGNTGGTA→AGNTGGTA
AGTAGAGA→AGTAGAGA
AGTAGCAA→AGTAGCAA
AGTAGCTA→AGTAGCTA
AGTAGGAA→AGTAGGAA
AGTAGGAG→AGTAGGAG
AGTAGGCA→AGTAGGCA
AGTAGGCG→AGTAGGCG
AGTAGGGA→AGTAGGGA
AGTAGGTA→AGTAGGTA
AGTAGGTG→AGTAGGTG
AGTAGGTT→AGTAGGTT
AGTAGTAA→AGTAGTAA
AGTAGTCA→AGTAGTCA
AGTAGTTA→AGTAGTTA
AGTATGGA→AGTAGG-A
AGTCAGCA→AGTCGC-A
AGTCAGTA→AGTCGT-A
AGTCCGTA→AGTCGT-A
AGTCGACA→AGTCGACA
AGTCGATA→AGTCGATA
AGTCGCTA→AGTCGCTA
AGTCGGAA→AGTCGGAA
AGTCGGCA→AGTCGGCA
AGTCGGCG→AGTCGGCG
AGTCGGGA→AGTCGGGA
AGTCGGTA→AGTCGGTA
AGTCGGTG→AGTCGGTG
AGTCGGTT→AGTCGGTT
AGTCTGTA→AGTCTGTA
AGTGAATA→AGT-GAAA
AGTGATTA→AGT-GATA
AGTGATTT→AGT-GATT
AGTGCCAA→AGT-GCCA
AGTGCGTA→AGT-GCGA
AGTGCTCA→AGT-GCTA
AGTGCTGA→AGT-GCTA
AGTGGAAA→AGT-GGAA
AGTGGAAC→AGT-GGAC
AGTGGAAT→AGT-GGAT
AGTGGACA→AGT-GGAA
AGTGGACG→AGT-GGAG
AGTGGAGA→AGT-GGAA
AGTGGAGC→AGT-GGAC
AGTGGAGT→AGT-GGAT
AGTGGATA→AGT-GGAA
AGTGGATC→AGT-GGAC
AGTGGATG→AGT-GGAG
AGTGGATT→AGT-GGAT
AGTGGCAA→AGT-GGCA
AGTGGCAC→AGT-GGCC
AGTGGCAG→AGT-GGCG
AGTGGCCA→AGT-GGCA
AGTGGCCC→AGT-GGCC
AGTGGCCG→AGT-GGCG
AGTGGCCT→AGT-GGCT
AGTGGCGA→AGT-GGCA
AGTGGCTA→AGT-GGCA
AGTGGCTC→AGT-GGCC
AGTGGCTG→AGT-GGCG
AGTGGCTT→AGT-GGCT
AGTGGGAA→AGTGGGAA
AGTGGGAT→AGTGGGAT
AGTGGGCA→AGTGGGCA
AGTGGGCT→AGTGGGCT
AGTGGGGA→AGTGGGGA
AGTGGGTA→AGTGGGTA
AGTGGTAA→AGT-GGTA
AGTGGTAG→AGT-GGTG
AGTGGTATAGT-GGTT
AGTGGTCA→AGT-GGTA
AGTGGTCG→AGT-GGTG
AGTGGTCT→AGT-GGTT
AGTGGTGT→AGT-GGTA
AGTGGTTA→AGT-GGTA
AGTGGTTC→AGT-GGTC
AGTGGTTG→AGT-GGTG
AGTGGTTT→AGT-GGTT
AGTGTCTA→AGT-GTCA
AGTGTTCA→AGT-GTTA
AGTGTTTA→AGT-GTTA
AGTTAGAA→AGTTAGAA
AGTTAGCA→AGTTAGCA
AGTTAGGA→AGTTGG-A
AGTTAGTA→AGTTAGTA
AGTTCGTA→AGTTCGTA
AGTTGATA→AGTTGATA
AGTTGCTA→AGTTGCTA
AGTTGGAA→AGTTGGAA
AGTTGGCA→AGTTGGCA
AGTTGGGA→AGTTGGGA
AGTTGGGT→AGTTGGGT
AGTTGGTA→AGTTGGTA
AGTTGGTC→AGTTGGTC
AGTTGGTG→AGTTGGTG
AGTTGGTT→AGTTGGTT
AGTTGTAA→AGTTGTAA
AGTTGTTA→AGTTGTTA
AGTTTGTA→AGTTTGTA
AGTYGGGA→AGTYGGGA
ATAAGGTA→ATAAGGTA
ATACGGAA→ATACGGAA
ATACGGTA→ATACGGTA
ATAGGTGA→ATA-GGTA
ATAGGTGG→ATA-GGTG
ATAGGTTA→ATA-GGTA
ATATGGAA→ATATGGAA
ATATGGAC→ATATGGAC
ATATGGAT→ATATGGAT
ATATGGTA→ATATGGTA
ATCAGGAA→ATCAGGAA
ATCAGGGA→ATCAGGGA
ATCAGGTA→ATCAGGTA
ATCAGTTA→ATCAGTTA
ATCCGGCA→ATCCGGCA
ATCCGGTA→ATCCGGTA
ATCGGAAA→ATC-GGAA
ATCGGATA→ATC-GGAA
ATCGGCAG→ATC-GGCG
ATCGGCCA→ATC-GGCA
ATCGGCGA→ATC-GGCA
ATCGGCTA→ATC-GGCA
ATCGGCTC→ATC-GGCC
ATCGGGAA→ATCGGGAA
ATCGGGTA→ATCGGGTA
ATCGGGTC→ATCGGGTC
ATCGGTCA→ATC-GGTA
ATCGGTGT→ATC-GGTT
ATCGGTTA→ATC-GGTA
ATCGGTTG→ATC-GGTG
ATCGGTTT→ATC-GGTT
ATCTGGCA→ATCTGGCA
ATCTGGCG→ATCTGGCG
ATCTGGGA→ATCTGGGA
ATCTGGTA→ATCTGGTA
ATGCGGTA→ATGCGGTA
ATGCGGTT→ATGCGGTT
ATGGGAGA→ATG-GGAA
ATGGGCTA→ATG-GGCA
ATGGGGTA→ATGGGGTA
ATGGGTTA→ATG-GGTA
ATGGGTTC→ATG-GGTC
ATGTGGAA→ATGTGGAA
ATGTGGAT→ATGTGGAT
ATGTGGTG→ATGTGGTG
ATTAGGAA→ATTAGGAA
ATTAGGTA→ATTAGGTA
ATTCGGAA→ATTCGGAA
ATTCGGTA→ATTCGGTA
ATTGCTCA→ATT-GCTA
ATTGGAGA→ATT-GGAA
ATTGGATA→ATT-GGAA
ATTGGCAC→ATT-GGCC
ATTGGCAG→ATT-GGCG
ATTGGCGA→ATT-GGCA
ATTGGCGT→ATT-GGCT
ATTGGCTA→ATT-GGCA
ATTGGGTA→ATT-GGGA
ATTGGTAA→ATT-GGTA
ATTGGTAT→ATT-GGTT
ATTGGTTA→ATT-GGTA
ATTGGTTC→ATT-GGTC
ATTTGGCA→ATTTGGCA
ATTTGGGA→ATTTGGGA
ATTTGGTA→ATTTGGTA
CAAAGGTA→CAAAGGTA
CAACAGCA→CAACAGCA
CAACGGCA→CAACGGCA
CAACGGCG→CAACGGCG
CAACGGCT→CAACGGCT
CAACGGGA→CAACGGGA
CAACGGTA→CAACGGTA
CAACGGTG→CAACGGTG
CAAGGGAG→CAAGGGAG
CAAGGGCA→CAAGGGCA
CAAGGTAA→CAA-GGTA
CAATGATA→CAATGATA
CAATGGCA→CAATGGCA
CAATGGGA→CAATGGGA
CAATGGGG→CAATGGGG
CAATGGTA→CAATGGTA
CAATGGTG→CAATGGTG
CACGGCGA→CAC-GGCA
CAGCGGCA→CAGCGGCA
CAGCGGGA→CAGCGGGA
CAGCGGTA→CAGCGGTA
CAGGGGTA→CAGGGGTA
CAGTGGTA→CAGTGGTA
CAGTGGTG→CAGTGGTG
CATAGGAG→CATAGGAG
CATGGCGA→CAT-GGCA
CATGGTAG→CAT-GGTG
CCCGGCGG→CCC-GGCG
CCCGGTGG→CCC-GGTG
CCTGGTGA→CCT-GGTA
CCTGGTGG→CCT-GGTG
CGATGGCA→CGATGGCA
CGATGGGG→CGATGGGG
CGCAGGTA→CGCAGGTA
CGCGGTCA→CGC-GGTA
CGTAGGTA→CGTAGGTA
CGTCGGTA→CGTCGGTA
CGTGGTTA→CGT-GGTA
CGTTGGCA→CGTTGGCA
CGTTGGTA→CGTTGGTA
CGTTTGTA→CGTTTGTA
CTCGGTAG→CTC-GGTG
CTGGGAGA→CTG-GGAA
CTGTGGAG→CTGTGGAG
CTGTGGTG→CTGTGGTG
CTTGGTGA→CTT-GGTA
CTTTGGCG→CTTTGGCG
GAACGGTA→GAACGGTA
GAAGGCTC→GAA-GGCC
GAAGGGAC→GAAGGGAC
GAAGGTTC→GAA-GGTC
GAATGGAC→GAATGGAC
GAATGGCA→GAATGGCA
GAATGGTA→GAATGGTA
GACAGGTA→GACAGGTA
GACGGACA→GAC-GGAA
GACGGCTA→GAC-GGCA
GACGGTTA→GAC-GGTA
GACGGTTC→GAC-GGTC
GACTGGAC→GACTGGAC
GACTGGCA→GACTGGCA
GACTGGTG→GACTGGTG
GAGAGGTT→GAGAGGTT
GAGCGGCA→GAGCGGCA
GAGCGGTA→GAGCGGTA
GAGGCCTA→GA--GGCA
GAGGGAAA→GAG-GGAA
GAGGGGTT→GAGGGGTT
GAGGGTTT→GAG-GGTT
GAGTGGAG→GAGTGGAG
GAGTGGTA→GAGTGGTA
GATCGGTC→GATCGGTC
GATGGACA→GAT-GGAA
GATGGATA→GAT-GGAA
GATGGCTG→GAT-GGCG

GATGGGAC→GAT-GGGC
GCCGGCAT→GCC-GGCT
GCCTGGTG→GCCTGGTG
GCTGGCGC→GCT-GGCC
GCTGGTGT→GCT-GGTT
GGCGGCAA→GGC-GGCA
GGGCGGAC→GGGCGGAC
GGGTGGTA→GGGTGGTA
GGTGGTTA→GGT-GGTA
GTCGGTTT→GTC-GGTT
GTTGGTGC→GTT-GGTC
TAACGGCA→TAACGGCA
TAATGGAA→TAATGGAA
TACTGGCT→TACTGGCT
TAGTGGTA→TAGTGGTA
TCCGGTGA→TCC-GGTA
TCTGGTGA→TCT-GGTA
TGCAGGCG→TGCAGGCG
TGGCGGCG→TGGCGGCG
TGTGGTTA→TGT-GGTA
TTATGGAT→TTATGGAT
TTTGGTAA→TTT-GGTA
GATGGTTG→GAT-GGTG
GCCGGTGA→GCC-GGTA
GCTCGGCG→GCTCGGCG
GCTGGGTA→GCT-GGGA
GCTGGTTG→GCT-GGTG
GGCGGCTA→GGC-GGCA
GGGGCTTT→GG--GGCT
GGTGGAAA→GGT-GGAA
GGTTGGCA→GGTTGGCA
GTGAGGAG→GTGAGGAG
GTTGGTTA→GTT-GGTA
TAACGGCT→TAACGGCT
TAATGGCA→TAATGGCA
TAGCGGTA→TAGCGGTA
TATCGGCA→TATCGGCA
TCCGGTGG→TCC-GGTG
TCTGGTGG→TCT-GGTG
TGCCGGCA→TGCCGGCA
TGTCGGCG→TGTCGGCG
TGTTGGCG→TGTTGGCG
TTCGGCAA→TTC-GGCA
TTTTGGTT→TTTTGGTT
GATGGTTT→GAT-GGTT
GCCGGTGC→GCC-GGTC
GCTGGATA→GCT-GGAA
GCTGGTAA→GCT-GGTA
GGAGGCAG→GGA-GGCG
GGCGGTCC→GGC-GGTC
GGGGGAAA→GGG-GGAA
GGTGGGTA→GGT-GGGA
GGTTGGTA→GGTTGGTA
GTGAGGGG→GTGAGGGG
NGTCGGTA→NGTCGGTA
TAACGGGA→TAACGGGA
TAATGGGA→TAATGGGA
TAGTGGAA→TAGTGGAA
TCAGTGGT→TCAGGG-T
TCGGGTGA→TCG-GGTA
TCTGGTGT→TCT-GGTT
TGCGGCAA→TGC-GGCA
TGTCGGTA→TGTCGGTA
TGTTGGTA→TGTTGGTA
TTGCGGAG→TTGCGGAG
AAAAGGACA→AAAAGGAA
AAACAGCAG→AAACAGCG
AAACGGCTA→AAACGGCA
AAACGGTTA→AAACGGTA
AAAGGCTAA→AAA-GGCA
AAAGGGTTA→AAAGGGTA
AAATAGGTA→AAATGGTA
AAATGGAGG→AAATGGAG
AAATGGCAT→AAATGGCT
AAATGGTAG→AAATGGTG
AAATTGGTA→AAATGGTA
AACAGCGCA→AACAGCGA
AACAGGCAA→AACAGGCA
AACAGGCTA→AACAGGCA
AACAGGTAT→AACAGGTT
AACAGTGAA→AACAGTGA
AACATGTCA→AACAGTCA
AACCGGAAA→AACCGGAA
AACCGGCAA→AACCGGCA
AACCGGGAG→AACCGGGG
AACCGGTGA→AACCGGTA
AACCTGGTC→AACCGGTC
AACGGACGA→AAC-GGAA
AACGGCAGG→AAC-GGCG
AACGGGATA→AACGGGAA
AACGGGTTA→AACGGGTA
AACTAGATA→AACTGATA
AACTCGGCA→AACTGGCA
AACTGGAAG→AACTGGAG
AACTGGATA→AACTGGAA
AACTGGCAT→AACTGGCT
AACTGGCTG→AACTGGCG
AACTGGTAA→AACTGGTA
AACTGGTCG→AACTGGTG
AACTGGTTG→AACTGGTG
AACTTGGGA→AACTGGGA
AAGAGGCAA→AAGAGGCA
AAGCAGGCA→AAGCGGCA
AAGCGGTAG→AAGCGGTG
AAGCGGTGA→AAGCGGTA
AAGCTGGGA→AAGCGGGA
AAGGGGTTA→AAGGGGTA
AAGTGGACA→AAGTGGAA
AAGTGGCAG→AAGTGGCG
AAGTGGTAT→AAGTGGTT
AAGTGGTTC→AAGTGGTC
AAGTTGGTA→AAGTGGTA
AATACGCAG→AATAGCAG
AATAGGACA→AATAGGAA
AATAGGCAG→AATAGGCG
AATAGGTAC→AATAGGTC
AATAGGTGA→AATAGGTA
GCCGCGGC→GCCGGG-C
GCCGGTGT→GCC-GGTT
GCTGGCCG→GCT-GGCG
GCTGGTGA→GCT-GGTA
GGAGGTCC→GGA-GGTC
GGCGGTCT→GGC-GGTT
GGGGTTTA→GG--GGTA
GGTGGTCC→GGT-GGTC
GGTTGGTC→GGTTGGTC
GTGGCTGA→GT--GGCA
RGGGGTTA→RGG-GGTA
TAACGGTA→TAACGGTA
TAATGGTA→TAATGGTA
TAGTGGCA→TAGTGGCA
TCCGGGCA→TCCGGGCA
TCTGGCGA→TCT-GGCA
TGATGGCG→TGATGGCG
TGCTGGCG→TGCTGGCG
TGTCGGTG→TGTCGGTG
TGTTGGTG→TGTTGGTG
TTTGGCAA→TTT-GGCA
AAAAGGATA→AAAAGGAA
AAACGGAGA→AAACGGAA
AAACGGGAG→AAACGGGG
AAACGTGTA→AAACGTGA
AAAGGCTAG→AAA-GGCG
AAAGGGTTC→AAAGGGTC
AAATGGAAA→AAATGGAA
AAATGGATA→AAATGGAA
AAATGGCCA→AAATGGCA
AAATGGTAT→AAATGGTT
AACAAGCAA→AACAGCAA
AACAGGAAG→AACAGGAG
AACAGGCAG→AACAGGCG
AACAGGCTG→AACAGGCG
AACAGGTCA→AACAGGTA
AACAGTTCA→AACAGTTA
AACCAGGCA→AACCGGCA
AACCGGAAG→AACCGGAG
AACCGGCAG→AACCGGCG
AACCGGTAA→AACCGGTA
AACCGGTTA→AACCGGTA
AACCTGGTG→AACCGGTG
AACGGATCA→AAC-GGAA
AACGGCTCA→AAC-GGCA
AACGGGCAG→AACGGGCG
AACGGTTAG→AAC-GGTG
AACTAGCAG→AACTGCAG
AACTGAATA→AACTGAAA
AACTGGAAT→AACTGGAT
AACTGGATT→AACTGGAT
AACTGGCCA→AACTGGCA
AACTGGGAA→AACTGGGA
AACTGGTAC→AACTGGTC
AACTGGTGA→AACTGGTA
AACTGGTTT→AACTGGTT
AACTTGGTA→AACTGGTA
AAGAGGCAG→AAGAGGCG
AAGCAGGTA→AAGCGGTA
AAGCGGTAT→AAGCGGTT
AAGCGGTGG→AAGCGGTG
AAGGGATCA→AAG-GGAA
AAGTAGTCA→AAGTGTCA
AAGTGGACG→AAGTGGAG
AAGTGGCTA→AAGTGGCA
AAGTGGTCA→AAGTGGTA
AAGTGGTTG→AAGTGGTG
AATAGGCAT→AATAGGCT
AATAGGTAG→AATAGGTG
AATAGGTTA→AATAGGTA
GCCGGAAC→GCC-GGAC
GCCTGGTA→GCCTGGTA
GCTGGCGA→GCT-GGCA
GCTGGTGC→GCT-GGTC
GGCGGATA→GGC-GGAA
GGCTGGGA→GGCTGGGA
GGGTGGGA→GGGTGGGA
GGTGGTCT→GGT-GGTT
GGTTGGTT→GGTTGGTT
GTTGGCAG→GTT-GGCG
TAAAGGTA→TAAAGGTA
TAAGGCAA→TAA-GGCA
TACCGGAA→TACCGGAA
TAGTGGCG→TAGTGGCG
TCCGGGGA→TCCGGGGA
TCTGGGAA→TCTGGGAA
TGATGGTT→TGATGGTT
TGCTGGGA→TGCTGGGA
TGTGGGAA→TGTGGGAA
TTACGGAT→TTACGGAT
TTTGGGAA→TTTGGGAA
AAAAGGTAG→AAAAGGTG
AAACGGATA→AAACGGAA
AAACGGTAA→AAACGGTA
AAAGGAATA→AAA-GGAA
AAAGGCTTA→AAA-GGCA
AAAGGTATA→AAA-GGTA
AAATGGAAG→AAATGGAG
AAATGGATT→AAATGGAT
AAATGGCTA→AAATGGCA
AAATGGTCA→AAATGGTA
AACAAGCCA→AACAGCCA
AACAGGACA→AACAGGAA
AACAGGCAT→AACAGGCT
AACAGGGAT→AACAGGGT
AACAGGTGA→AACAGGTA
AACATGGAA→AACAGGAA
AACCCGATA→AACCGATA
AACCGGACA→AACCGGAA
AACCGGCAT→AACCGGCT
AACCGGTAG→AACCGGTG
AACCGGTTG→AACCGGTG
AACGAGGCA→AAC-GAGA
AACGGATGA→AAC-GGAA
AACGGCTGA→AAC-GGCA
AACGGGGAT→AACGGGGT
AACGGTTCA→AAC-GGTA
AACTAGGCA→AACTGGCA
AACTGAGTA→AACTGAGA
AACTGGACA→AACTGGAA
AACTGGCAA→AACTGGCA
AACTGGCCG→AACTGGCG
AACTGGGAG→AACTGGGG
AACTGGTAG→AACTGGTG
AACTGGTGT→AACTGGTT
AACTGTATA→AACTGTAA
AAGAAGGCA→AAGAGGCA
AAGAGGTAG→AAGAGGTG
AAGCGGATA→AAGCGGAA
AAGCGGTCA→AAGCGGTA
AAGCGGTTA→AAGCGGTA
AAGGGATTA→AAG-GGAA
AAGTCCGGC→AAGTGG-C
AAGTGGAGA→AAGTGGAA
AAGTGGTAA→AAGTGGTA
AAGTGGTCG→AAGTGGTG
AAGTTCGGC→AAGTGG-C
AATAGGAAA→AATAGGAA
AATAGGATG→AATAGGAG
AATAGGCTA→AATAGGCA
AATAGGTAT→AATAGGTT
AATAGGTTC→AATAGGTC

Chapter V – Appendix

AATAGTGCA→AATAGTGA
AATCAGGCA→AATCGGCA
AATCCGGCC→AATCGGCC
AATCGGAAG→AATCGGAG
AATCGGCAG→AATCGGCG
AATCGGGCA→AATCGGGA
AATCGGTCA→AATCGGTA
AATCGTGTA→AATCGTGA
AATCTGGTA→AATCGGTA
AATGGATCA→AAT-GGAA
AATGGCGAG→AAT-GGCG
AATGGCTGA→AAT-GGCA
AATGGGCAT→AATGGGCT
AATGGGTGA→AATGGGTA
AATGGTTAA→AAT-GGTA
AATGGTTTA→AAT-GGTA
AATTAGGTA→AATTGGTA
AATTCGGCA→AATTGGCA
AATTGGAAC→AATTGGAC
AATTGGATA→AATTGGAA
AATTGGCAG→AATTGGCG
AATTGGCTA→AATTGGCA
AATTGGGAT→AATTGGGT
AATTGGTAG→AATTGGTG
AATTGGTGA→AATTGGTA
AATTGTTAG→AATTGTTG
AATTTGGTC→AATTGGTC
ACAAGGTTA→ACAAGGTA
ACATGGTTA→ACATGGTA
ACCCGGATA→ACCCGGAA
ACCGGCTGT→ACC-GGCT
ACCTGGCCA→ACCTGGCA
ACGACGGTA→ACGAGGTA
ACGTGGCTA→ACGTGGCA
ACTCGGATA→ACTCGGAA
ACTGGGTGT→ACTGGGTT
ACTTGGTAA→ACTTGGTA
AGAAAGGCA→AGAAGGCA
AGAAGGTCA→AGAAGGTA
AGAATGGCA→AGAAGGCA
AGACGGATA→AGACGGAA
AGACGGGGA→AGACGGGA
AGACTGGGA→AGACGGGA
AGAGATCTA→AGA-GATA
AGAGGATCA→AGA-GGAA
AGAGGCCCA→AGA-GGCA
AGAGGCGAG→AGA-GGCG
AGAGGGCTA→AGAGGGCA
AGAGGTCGT→AGA-GGTT
AGAGGTTGA→AGA-GGTA
AGATCGGTA→AGATGGTA
AGATGGAGT→AGATGGAT
AGATGGCGA→AGATGGCA
AGATGGTCA→AGATGGTA
AGATGGTTG→AGATGGTG
AGCACGGAC→AGCAGGAC
AGCAGAGTA→AGCAGAGA
AGCAGGATA→AGCAGGAA
AGCAGGCTA→AGCAGGCA
AGCAGGTCT→AGCAGGTT
AGCAGTCAA→AGCAGTCA
AGCCAGACA→AGCCGACA
AGCCAGGTG→AGCCGGTG
AGCCCGGGA→AGCCGGGA
AGCCGGAAA→AGCCGGAA
AGCCGGCCA→AGCCGGCA
AGCCGGTAT→AGCCGGTT
AGCCTAGTA→AGCCAGTA
AGCCTGGCC→AGCCGGCC
AGCCTGGTG→AGCCGGTG
AGCGACTTA→AGC-GACA
AGCGGACCA→AGC-GGAA
AATATGGCA→AATAGGCA
AATCAGGTA→AATCGGTA
AATCCGGTA→AATCGGTA
AATCGGACA→AATCGGAA
AATCGGCAT→AATCGGCT
AATCGGGTA→AATCGGGA
AATCGGTCG→AATCGGTG
AATCGTTAG→AATCGTTG
AATGGAAGA→AAT-GGAA
AATGGATTA→AAT-GGAA
AATGGCTAA→AAT-GGCA
AATGGCTTA→AAT-GGCA
AATGGGGAG→AATGGGGG
AATGGGTTA→AATGGGTA
AATGGTTAG→AAT-GGTG
AATGGTTTC→AAT-GGTC
AATTAGGTG→AATTGGTG
AATTCGGTA→AATTGGTA
AATTGGAAG→AATTGGAG
AATTGGATC→AATTGGAC
AATTGGCAT→AATTGGCT
AATTGGCTG→AATTGGCG
AATTGGGTA→AATTGGGA
AATTGGTAT→AATTGGTT
AATTGGTTA→AATTGGTA
AATTTGGAC→AATTGGAC
AATTTGGTT→AATTGGTT
AATTTGGTC→AATTGGTC
AGAAGAGTA→AGAAGAGA
AGAAGGTCT→AGAAGGTT
AGACAGGTA→AGACGGTA
AGACGGCAA→AGACGGCA
AGACGGTCA→AGACGGTA
AGACTGGTA→AGACGGTA
AGAGGACAA→AGA-GGAA
AGAGGATGA→AGA-GGAA
AGAGGCCGA→AGA-GGCA
AGAGGCTAA→AGA-GGCA
AGAGGGTTC→AGAGGGTC
AGAGGTCTA→AGA-GGTA
AGAGGTTTA→AGA-GGTA
AGATGAGTA→AGATGAGA
AGATGGATA→AGATGGAA
AGATGGCTA→AGATGGCA
AGATGGTCT→AGATGGTT
AGATTGGGA→AGATGGGA
AGCACGGTA→AGCAGGTA
AGCAGCCCA→AGCAGCCA
AGCAGGCAA→AGCAGGCA
AGCAGGCTG→AGCAGGCG
AGCAGGTGA→AGCAGGTA
AGCATGGAA→AGCAGGAA
AGCCAGGAA→AGCCGGAA
AGCCCGATA→AGCCGATA
AGCCCGGTA→AGCCGGTA
AGCCGGACA→AGCCGGAA
AGCCGGCTA→AGCCGGCA
AGCCGGTCA→AGCCGGTA
AGCCTGATA→AGCCGATA
AGCCTGGCG→AGCCGGCG
AGCCTGGTT→AGCCGGTT
AGCGATCTA→AGC-GATA
AGCGGACGA→AGC-GGAA
AATATGGTA→AATAGGTA
AATCAGGTT→AATCGGTT
AATCGATTA→AATCGATA
AATCGGATA→AATCGGAA
AATCGGCCA→AATCGGCA
AATCGGTAA→AATCGGTA
AATCGGTGA→AATCGGTA
AATCTGGAA→AATCGGAA
AATGGAATA→AAT-GGAA
AATGGCAAA→AAT-GGCA
AATGGCTAG→AAT-GGCG
AATGGGAAG→AATGGGAG
AATGGGTAA→AATGGGTA
AATGGGTTG→AATGGGTG
AATGGTTAT→AAT-GGTT
AATGTGGCA→AATGGGCA
AATTAGTAG→AATTGTAG
AATTGAATA→AATTGAAA
AATTGGAAT→AATTGGAT
AATTGGATT→AATTGGAT
AATTGGCCA→AATTGGCA
AATTGGCTT→AATTGGCT
AATTGGNTA→AATTGGNA
AATTGGTCA→AATTGGTA
AATTGGTTG→AATTGGTG
AATTTGGCA→AATTGGCA
AATTTGTAG→AATTGTAG
ACACGGTTC→ACACGGTC
ACCAAGCAA→ACCAGCAA
ACCCGGTGG→ACCCGGTG
ACCGGTGGT→ACC-GGTT
ACCTGGTGA→ACCTGGTA
ACGCGGTGT→ACGCGGTT
ACGTGGTGT→ACGTGGTT
ACTCGGTAG→ACTCGGTG
ACTTGGAAA→ACTTGGAA
ACTTGGTGA→ACTTGGTA
AGAAGCGCA→AGAAGCGA
AGAAGGTGA→AGAAGGTA
AGACCGGCA→AGACGGCA
AGACGGCAT→AGACGGCT
AGACGGTGA→AGACGGTA
AGAGACCCA→AGA-GACA
AGAGGACCA→AGA-GGAA
AGAGGATTA→AGA-GGAA
AGAGGCCGT→AGA-GGCT
AGAGGCTCA→AGA-GGCA
AGAGGTCAA→AGA-GGTA
AGAGGTGAG→AGA-GGTG
AGAGTCCTA→AGA-GTCA
AGATGGAAA→AGATGGAA
AGATGGCAA→AGATGGCA
AGATGGGCA→AGATGGGA
AGATGGTGA→AGATGGTA
AGATTGGTA→AGATGGTA
AGCACGGTC→AGCAGGTC
AGCAGCCGA→AGCAGCCA
AGCAGGCAG→AGCAGGCG
AGCAGGGGA→AGCAGGGA
AGCAGGTTA→AGCAGGTA
AGCATGGCA→AGCAGGCA
AGCCAGGCA→AGCCGGCA
AGCCCGGAA→AGCCGGAA
AGCCCGGTC→AGCCGGTC
AGCCGGAGA→AGCCGGAA
AGCCGGGAT→AGCCGGGT
AGCCGGTGA→AGCCGGTA
AGCCTGGAA→AGCCGGAA
AGCCTGGCT→AGCCGGCT
AGCCTTGTA→AGCCTGTA
AGCGCGGAA→AGCGGGAA
AGCGGACTA→AGC-GGAA
AATCAGACA→AATCGACA
AATCCGGAA→AATCGGAA
AATCGCGCA→AATCGCGA
AATCGGCAA→AATCGGCA
AATCGGCTA→AATCGGCA
AATCGGTAG→AATCGGTG
AATCGGTTA→AATCGGTA
AATCTGGCA→AATCGGCA
AATGGACCA→AAT-GGAA
AATGGCAAG→AAT-GGCG
AATGGCTAT→AAT-GGCT
AATGGGATA→AATGGGAA
AATGGGTAG→AATGGGTG
AATGGTAGG→AAT-GGTG
AATGGTTCA→AAT-GGTA
AATTAGGCA→AATTGGCA
AATTCGATA→AATTGATA
AATTGATAG→AATTGATG
AATTGGACA→AATTGGAA
AATTGGCAA→AATTGGCA
AATTGGCCT→AATTGGCT
AATTGGGAA→AATTGGGA
AATTGGTAA→AATTGGTA
AATTGGTCG→AATTGGTG
AATTGGTTT→AATTGGTT
AATTTGGGA→AATTGGGA
ACAAGGTAA→ACAAGGTA
ACACGGTTG→ACACGGTG
ACCAGGCGA→ACCAGGCA
ACCCGGTTT→ACCCGGTT
ACCTGGACA→ACCTGGAA
ACCTGGTGG→ACCTGGTG
ACGGTGGTT→ACGGGGTT
ACGTGGTTA→ACGTGGTA
ACTCGGTTA→ACTCGGTA
ACTTGGACA→ACTTGGAA
ACTTGGTGT→ACTTGGTT
AGAAGGATA→AGAAGGAA
AGAAGGTTA→AGAAGGTA
AGACCGGGA→AGACGGGA
AGACGGCCA→AGACGGCA
AGACGGTTA→AGACGGTA
AGAGACCTA→AGA-GACA
AGAGGACGA→AGA-GGAA
AGAGGCACG→AGA-GGCG
AGAGGCCTA→AGA-GGCA
AGAGGCTGA→AGA-GGCA
AGAGGTCCA→AGA-GGTA
AGAGGTTAA→AGA-GGTA
AGATAGGTA→AGATGGTA
AGATGGACA→AGATGGAA
AGATGGCAT→AGATGGCT
AGATGGTAA→AGATGGTA
AGATGGTTA→AGATGGTA
AGCAAGGAA→AGCAGGAA
AGCACGTTA→AGCAGTTA
AGCAGGAAA→AGCAGGAA
AGCAGGCCA→AGCAGGCA
AGCAGGGTA→AGCAGGGA
AGCAGGTTC→AGCAGGTC
AGCATGGGG→AGCAGGGG
AGCCAGGGA→AGCCGGGA
AGCCCGGCA→AGCCGGCA
AGCCCGGTG→AGCCGGTG
AGCCGGATA→AGCCGGAA
AGCCGGGCA→AGCCGGGA
AGCCGGTTA→AGCCGGTA
AGCCTGGAT→AGCCGGAT
AGCCTGGGA→AGCCGGGA
AGCGACCAA→AGC-GACA
AGCGGAATA→AGC-GGAA
AGCGGACTT→AGC-GGAT
AATCAGCCA→AATCGCCA
AATCCGGCA→AATCGGCA
AATCGGAAA→AATCGGAA
AATCGGCAC→AATCGGCC
AATCGGGAG→AATCGGGG
AATCGGTAT→AATCGGTT
AATCGGTTG→AATCGGTG
AATCTGGGA→AATCGGGA
AATGGATAA→AAT-GGAA
AATGGCAGG→AAT-GGCG
AATGGCTCA→AAT-GGCA
AATGGGCAG→AATGGGCG
AATGGGTCA→AATGGGTA
AATGGTCAA→AAT-GGTA
AATGGTTGA→AAT-GGTA
AATTAGGCG→AATTGGCG
AATTCGCAG→AATTGCAG
AATTGGAAA→AATTGGAA
AATTGGAGA→AATTGGAA
AATTGGCAC→AATTGGCC
AATTGGCGA→AATTGGCA
AATTGGGAG→AATTGGGG
AATTGGTAC→AATTGGTC
AATTGGTCT→AATTGGTT
AATTGTATA→AATTGTAA
AATTTGGTA→AATTGGTA
ACAAGGTGT→ACAAGGTT
ACAGGTTGT→ACA-GGTT
ACCAGGTTA→ACCAGGTA
ACCGGCGCG→ACC-GGCG
ACCTGGATA→ACCTGGAA
ACCTGGTTA→ACCTGGTA
ACGTGGATA→ACGTGGAA
ACGTTGGCT→ACGTGGCT
ACTGGACTA→ACT-GGAA
ACTTGGCTA→ACTTGGCA
ACTTGGTTA→ACTTGGTA
AGAAGGTAA→AGAAGGTA
AGAAGGTTC→AGAAGGTC
AGACCGGTA→AGACGGTA
AGACGGCTA→AGACGGCA
AGACTGGCA→AGACGGCA
AGAGACTTA→AGA-GACA
AGAGGACTA→AGA-GGAA
AGAGGCCAA→AGA-GGCA
AGAGGCCTT→AGA-GGCT
AGAGGCTTA→AGA-GGCA
AGAGGTCGA→AGA-GGTA
AGAGGTTCA→AGA-GGTA
AGATAGTCT→AGATGTCT
AGATGGAGA→AGATGGAA
AGATGGCCA→AGATGGCA
AGATGGTAG→AGATGGTG
AGATGGTTC→AGATGGTC
AGCAAGGTA→AGCAGGTA
AGCAGACGC→AGCAGACC
AGCAGGACA→AGCAGGAA
AGCAGGCGA→AGCAGGCA
AGCAGGTCA→AGCAGGTA
AGCAGGTTG→AGCAGGTG
AGCATGGTA→AGCAGGTA
AGCCAGGTA→AGCCGGTA
AGCCCGGCG→AGCCGGCG
AGCCGCATA→AGCCGCAA
AGCCGGCAT→AGCCGGCT
AGCCGGGTA→AGCCGGGA
AGCCTAATA→AGCCTAAA
AGCCTGGCA→AGCCGGCA
AGCCTGGTA→AGCCGGTA
AGCGACCTA→AGC-GACA
AGCGGACAA→AGC-GGAA
AGCGGAGTA→AGC-GGAA

AGCGGATAA→AGC-GGAA
AGCGGCCAA→AGC-GGCA
AGCGGCGAG→AGC-GGCG
AGCGGCTTA→AGC-GGCA
AGCGGGATA→AGCGGGAA
AGCGGGTGA→AGCGGGTA
AGCGGTCCA→AGC-GGTA
AGCGGTTAG→AGC-GGTG
AGCGTGGGA→AGCGGGGA
AGCTAGGAA→AGCTGGAA
AGCTCGGAA→AGCTGGAA
AGCTCGTTA→AGCTGTTA
AGCTGGAAC→AGCTGGAC
AGCTGGATG→AGCTGGAG
AGCTGGCGA→AGCTGGCA
AGCTGGGAG→AGCTGGGG
AGCTGGTAC→AGCTGGTC
AGCTGGTCT→AGCTGGTT
AGCTGGTTA→AGCTGGTA
AGCTTAGTA→AGCTAGTA
AGCTTGGKA→AGCTGGKA
AGGAGGATA→AGGAGGAA
AGGAGGTTC→AGGAGGTC
AGGCGGATA→AGGCGGAA
AGGCGGTTA→AGGCGGTA
AGGGGACCA→AGG-GGAA
AGGGGCCAA→AGG-GGCA
AGGGGCTTA→AGG-GGCA
AGGGGTCGA→AGG-GGTA
AGGGGTTTA→AGG-GGTA
AGGTGGAAA→AGGTGGAA
AGGTGGCAA→AGGTGGCA
AGGTGGTAT→AGGTGGTT
AGGTGGTTA→AGGTGGTA
AGGTTGTTA→AGGTGTTA
AGNTGGTTA→AGNTGGTA
AGTACGGTA→AGTAGGTA
AGTAGGAGA→AGTAGGAA
AGTAGGTAA→AGTAGGTA
AGTAGTCGA→AGTAGTCA
AGTCAGGCA→AGTCGGCA
AGTCCGGCA→AGTCGGCA
AGTCGAATG→AGTCGAAG
AGTCGGATA→AGTCGGAA
AGTCGGCCG→AGTCGGCG
AGTCGGTAC→AGTCGGTC
AGTCGGTTA→AGTCGGTA
AGTCTGGCA→AGTCGGCA
AGTGAACGA→AGT-GAAA
AGTGACTGA→AGT-GACA
AGTGCCTAA→AGT-GCCA
AGTGGACGA→AGT-GGAA
AGTGGATGA→AGT-GGAA
AGTGGCATA→AGT-GGCA
AGTGGCCTA→AGT-GGCA
AGTGGCTCA→AGT-GGCA
AGTGGGATA→AGTGGGAA
AGTGGGTCA→AGTGGGTA
AGTGGTCAT→AGT-GGTT
AGTGGTTAA→AGT-GGTA
AGTGGTTTT→AGT-GGTT
AGTTAGGGA→AGTTGGGA
AGTTCGGCA→AGTTGGCA
AGTTCGTTA→AGTTGTTA
AGTTGGAAA→AGTTGGAA
AGTTGGAGA→AGTTGGAA
AGTTGGCAC→AGTTGGCC
AGTTGGCGA→AGTTGGCA
AGTTGGTAA→AGTTGGTA
AGTTGGTCG→AGTTGGTG
AGTTGGTTC→AGTTGGTC
AGTTGTTTA→AGTTGTTA

AGCGGATCA→AGC-GGAA
AGCGGCCAG→AGC-GGCG
AGCGGCTAA→AGC-GGCA
AGCGGGAAA→AGCGGGAA
AGCGGGCAC→AGCGGGCC
AGCGGGTTA→AGCGGGTA
AGCGGTCGA→AGC-GGTA
AGCGGTTCA→AGC-GGTA
AGCGTGGTA→AGCGGGTA
AGCTAGGCA→AGCTGGCA
AGCTCGGCA→AGCTGGCA
AGCTGAATA→AGCTGAAA
AGCTGGACA→AGCTGGAA
AGCTGGCAA→AGCTGGCA
AGCTGGCTA→AGCTGGCA
AGCTGGGAT→AGCTGGGT
AGCTGGTAG→AGCTGGTG
AGCTGGTGA→AGCTGGTA
AGCTGGTTG→AGCTGGTG
AGCTTCGTA→AGCTCGTA
AGCTTGGTA→AGCTGGTA
AGGAGGCTA→AGGAGGCA
AGGCAGGTA→AGGCGGTA
AGGCGGCTA→AGGCGGCA
AGGCTGGCA→AGGCGGCA
AGGGGACTA→AGG-GGAA
AGGGGCCCA→AGG-GGCA
AGGGGGGTA→AGGGGGGA
AGGGGTCTA→AGG-GGTA
AGGTAGGCA→AGGTGGCA
AGGTGGACA→AGGTGGAA
AGGTGGCCA→AGGTGGCA
AGGTGGTCA→AGGTGGTA
AGGTGGTTC→AGGTGGTC
AGNCGGTTA→AGNCGGTA
AGTAAGATA→AGTAGATA
AGTAGACGA→AGTAGACA
AGTAGGATA→AGTAGGAA
AGTAGGTAG→AGTAGGTG
AGTATGGCA→AGTAGGCA
AGTCAGGGA→AGTCGGGA
AGTCCGGCG→AGTCGGCG
AGTCGCGTA→AGTCGCGA
AGTCGGCAA→AGTCGGCA
AGTCGGCTA→AGTCGGCA
AGTCGGTAG→AGTCGGTG
AGTCGGTTG→AGTCGGTG
AGTCTGGGA→AGTCGGGA
AGTGACCGA→AGT-GACA
AGTGACTTA→AGT-GACA
AGTGGAACA→AGT-GGAA
AGTGGACTA→AGT-GGAA
AGTGGATGG→AGT-GGAG
AGTGGCCAA→AGT-GGCA
AGTGGCCTC→AGT-GGCC
AGTGGCTGA→AGT-GGCA
AGTGGGCAA→AGTGGGCA
AGTGGGTTA→AGTGGGTA
AGTGGTCCA→AGT-GGTA
AGTGGTTAG→AGT-GGTG
AGTGTGGTA→AGTGGGTA
AGTTAGGTA→AGTTGGTA
AGTTCGGGA→AGTTGGGA
AGTTGAATA→AGTTGAAA
AGTTGGAAC→AGTTGGAC
AGTTGGAGT→AGTTGGAT
AGTTGGCAG→AGTTGGCG
AGTTGGCTA→AGTTGGCA
AGTTGGTAC→AGTTGGTC
AGTTGGTCT→AGTTGGTT
AGTTGGTTG→AGTTGGTG
AGTTTGGAA→AGTTGGAA

AGCGGATGA→AGC-GGAA
AGCGGCCCA→AGC-GGCA
AGCGGCTAG→AGC-GGCG
AGCGGGACA→AGCGGGAA
AGCGGGCTA→AGCGGGCA
AGCGGTAAG→AGC-GGTG
AGCGGTCTA→AGC-GGTA
AGCGGTTGA→AGC-GGTA
AGCGTTTGA→AGC-GTTA
AGCTAGGTA→AGCTGGTA
AGCTCGGGA→AGCTGGGA
AGCTGCATA→AGCTGCAA
AGCTGGAGA→AGCTGGAA
AGCTGGCAG→AGCTGGCG
AGCTGGCTG→AGCTGGCG
AGCTGGGGA→AGCTGGGA
AGCTGGTAT→AGCTGGTT
AGCTGGTGC→AGCTGGTC
AGCTGGTTT→AGCTGGTT
AGCTTGGAA→AGCTGGAA
AGCTTGGTG→AGCTGGTG
AGGAGGTCA→AGGAGGTA
AGGCCGGCA→AGGCGGCA
AGGCGGTAA→AGGCGGTA
AGGCTGGGA→AGGCGGGA
AGGGGATAA→AGG-GGAA
AGGGGCCGA→AGG-GGCA
AGGGGGTTA→AGGGGGTA
AGGGGTTAA→AGG-GGTA
AGGTAGGTA→AGGTGGTA
AGGTGGAGT→AGGTGGAT
AGGTGGCTA→AGGTGGCA
AGGTGGTCC→AGGTGGTC
AGGTGGTTG→AGGTGGTG
AGNGGCCAA→AGN-GGCA
AGTAGAGTA→AGTAGAGA
AGTAGGCAT→AGTAGGCT
AGTAGGTCA→AGTAGGTA
AGTATGGGA→AGTAGGGA
AGTCAGGTA→AGTCGGTA
AGTCCGGGA→AGTCGGGA
AGTCGGAAA→AGTCGGAA
AGTCGGCAG→AGTCGGCG
AGTCGGGCA→AGTCGGGA
AGTCGGTAT→AGTCGGTT
AGTCGTGTA→AGTCGTGA
AGTCTGGNA→AGTCGGNA
AGTGACCTA→AGT-GACA
AGTGATCGA→AGT-GATA
AGTGGAATA→AGT-GGAA
AGTGGAGCA→AGT-GGAA
AGTGGATTA→AGT-GGAA
AGTGGCCAG→AGT-GGCG
AGTGGCGCT→AGT-GGCT
AGTGGCTTA→AGT-GGCA
AGTGGGCCA→AGTGGGCA
AGTGGNCTA→AGT-GGNA
AGTGGTCGA→AGT-GGTA
AGTGGTTCA→AGT-GGTA
AGTNGGCTA→AGTNGGCA
AGTTAGTCA→AGTTGTCA
AGTTCGGTA→AGTTGGTA
AGTTGATAG→AGTTGATG
AGTTGGAAG→AGTTGGAG
AGTTGGATA→AGTTGGAA
AGTTGGCAT→AGTTGGCT
AGTTGGGAA→AGTTGGGA
AGTTGGTAG→AGTTGGTG
AGTTGGTGA→AGTTGGTA
AGTTTGGCA→AGTTGGCA

AGCGGATTA→AGC-GGAA
AGCGGCCGA→AGC-GGCA
AGCGGCTCA→AGC-GGCA
AGCGGGAGA→AGCGGGAA
AGCGGGGAG→AGCGGGGG
AGCGGTATA→AGC-GGTA
AGCGGTCTC→AGC-GGTC
AGCGGTTTA→AGC-GGTA
AGCMTGGTA→AGCMGGTA
AGCTAGGTG→AGCTGGTG
AGCTCGGTA→AGCTGGTA
AGCTGCCAA→AGCTGCCA
AGCTGGAGT→AGCTGGAT
AGCTGGCAT→AGCTGGCT
AGCTGGGAA→AGCTGGAA
AGCTGGGTA→AGCTGGGA
AGCTGGTCA→AGCTGGTA
AGCTGGTGG→AGCTGGTG
AGCTGTATA→AGCTGTAA
AGCTTGGCA→AGCTGGCA
AGGAGGAAA→AGGAGGAA
AGGAGGTGA→AGGAGGTA
AGGCCGGTA→AGGCGGTA
AGGCGGTCA→AGGCGGTA
AGGCTGGTA→AGGCGGTA
AGGGGATCA→AGG-GGAA
AGGGGCCTA→AGG-GGCA
AGGGGTCAA→AGG-GGTA
AGGGGTTCA→AGG-GGTA
AGGTCGGTA→AGGTGGTA
AGGTGGATA→AGGTGGAA
AGGTGGTAA→AGGTGGTA
AGGTGGTGA→AGGTGGTA
AGGTGGTTT→AGGTGGTT
AGNGGCTGA→AGN-GGCA
AGTAAGGTA→AGTAGGTA
AGTAGGAAA→AGTAGGAA
AGTAGGCCA→AGTAGGCA
AGTAGGTGA→AGTAGGTA
AGTATGGTA→AGTAGGTA
AGTCAGGTT→AGTCGGTT
AGTCCGGTA→AGTCGGTA
AGTCGGACA→AGTCGGAA
AGTCGGCAT→AGTCGGCT
AGTCGGGTA→AGTCGGGA
AGTCGGTCA→AGTCGGTA
AGTCTGGAA→AGTCGGAA
AGTCTGGTA→AGTCGGTA
AGTGACCTG→AGT-GACG
AGTGATCTA→AGT-GATA
AGTGGACAA→AGT-GGAA
AGTGGATAA→AGT-GGAA
AGTGGCACA→AGT-GGCA
AGTGGCCCA→AGT-GGCA
AGTGGCGTA→AGT-GGCA
AGTGGCTTC→AGT-GGCC
AGTGGGCGA→AGTGGGCA
AGTGGTAAG→AGT-GGTG
AGTGGTCTA→AGT-GGTA
AGTGGTTGA→AGT-GGTA
AGTNGGTTA→AGTNGGTA
AGTTAGTTA→AGTTGTTA
AGTTCGGTC→AGTTGGTC
AGTTGATTA→AGTTGATA
AGTTGGAAT→AGTTGGAT
AGTTGGATT→AGTTGGAT
AGTTGGCCA→AGTTGGCA
AGTTGGGAT→AGTTGGGT
AGTTGGTAT→AGTTGGTT
AGTTGGTGG→AGTTGGTG
AGTTGTATA→AGTTGTAA
AGTTTGGGA→AGTTGGGA

AGCGGCATA→AGC-GGCA
AGCGGCCTA→AGC-GGCA
AGCGGCTGA→AGC-GGCA
AGCGGGAGT→AGCGGGAT
AGCGGGTCA→AGCGGGTA
AGCGGTCAA→AGC-GGTA
AGCGGTTAA→AGC-GGTA
AGCGRCCAA→AGC-GRCA
AGCNCGGTA→AGCNGGTA
AGCTAGTTA→AGCTGTTA
AGCTCGGTT→AGCTGGTT
AGCTGGAAA→AGCTGGAA
AGCTGGATA→AGCTGGAA
AGCTGGCCA→AGCTGGCA
AGCTGGGAC→AGCTGGGC
AGCTGGTAA→AGCTGGTA
AGCTGGTCG→AGCTGGTG
AGCTGGTGT→AGCTGGTT
AGCTNGGTA→AGCTGGTA
AGCTTGGGA→AGCTGGGA
AGGAGGACA→AGGAGGAA
AGGAGGTTA→AGGAGGTA
AGGCGGACA→AGGCGGAA
AGGCGGTCC→AGGCGGTC
AGGGGACAA→AGG-GGAA
AGGGGATTA→AGG-GGAA
AGGGGCTGA→AGG-GGCA
AGGGGTCCA→AGG-GGTA
AGGGGTTGA→AGG-GGTA
AGGTGATTA→AGGTGATA
AGGTGGATT→AGGTGGAT
AGGTGGTAG→AGGTGGTG
AGGTGGTGG→AGGTGGTG
AGGTTGGTA→AGGTGGTA
AGNTAGGTA→AGNTGGTA
AGTACGGAA→AGTAGGAA
AGTAGGACA→AGTAGGAA
AGTAGGCTA→AGTAGGCA
AGTAGGTTA→AGTAGGTA
AGTCAGGAA→AGTCGGAA
AGTCCGGAA→AGTCGGAA
AGTCCGGTG→AGTCGGTG
AGTCGGAGT→AGTCGGAT
AGTCGGCCA→AGTCGGCA
AGTCGGTAA→AGTCGGTA
AGTCGGTGA→AGTCGGTA
AGTCTGGAC→AGTCGGAC
AGTCTGGTG→AGTCGGTG
AGTGACTCA→AGT-GACA
AGTGATTTA→AGT-GATA
AGTGGACCA→AGT-GGAA
AGTGGATCA→AGT-GGAA
AGTGGCAGC→AGT-GGCC
AGTGGCCGA→AGT-GGCA
AGTGGCTAA→AGT-GGCA
AGTGGCTTT→AGT-GGCT
AGTGGGGAA→AGTGGGAA
AGTGGTCAA→AGT-GGTA
AGTGGTCTT→AGT-GGTT
AGTGGTTTA→AGT-GGTA
AGTTAGGCA→AGTTGGCA
AGTTCGCCA→AGTTGCCA
AGTTCGGTT→AGTTGGTT
AGTTGCGTA→AGTTGCGA
AGTTGGACA→AGTTGGAA
AGTTGGCAA→AGTTGGCA
AGTTGGCCG→AGTTGGCG
AGTTGGGTA→AGTTGGGA
AGTTGGTCA→AGTTGGTA
AGTTGGTTA→AGTTGGTA
AGTTGTGTA→AGTTGTGA
AGTTTGGTA→AGTTGGTA

Chapter V – Appendix

AGTTTGGTG→AGTTGGTG
ANCGGCCCA→ANC-GGCA
ATACGGTTA→ATACGGTA
ATCAGGAAA→ATCAGGAA
ATCAGGTCA→ATCAGGTA
ATCCGGATA→ATCCGGAA
ATCCGGTGA→ATCCGGTA
ATCGGGAAA→ATCGGGAA
ATCGGTTCA→ATC-GGTA
ATCRGGAAA→ATCRGGAA
ATCTGGCCA→ATCTGGCA
ATCTGGTGG→ATCTGGTG
ATGAGGCAC→ATGAGGCC
ATGTGGAAA→ATGTGGAA
ATTAGGTAT→ATTAGGTT
ATTCGGTTA→ATTCGGTA
ATTGGTTTT→ATT-GGTT
ATTTGGTTA→ATTTGGTA
CAACGGATA→CAACGGAA
CAACTGGCA→CAACGGCA
CAATCGGTC→CAATGGTC
CAATGGTTG→CAATGGTG
CACTGGCGA→CACTGGCA
CAGCGGTAA→CAGCGGTA
CAGCTGGGA→CAGCGGGA
CAGTCGGTA→CAGTGGTA
CAGTTGGTG→CAGTGGTG
CCGGGGTGG→CCGGGGTG
CGCTGGCGA→CGCTGGCA
CGGTGGCAG→CGGTGGCG
CGTTTGATA→CGTTGATA
CTCTGGTGG→CTCTGGTG
GAAATGGTA→GAAAGGTA
GAATGGAGC→GAATGGAC
GAATGGTTA→GAATGGTA
GACCCGGTG→GACCGGTG
GACTGGTGT→GACTGGTT
GAGTGGCAA→GAGTGGCA
GATGGATTC→GAT-GGAC
GATTGGAAA→GATTGGAA
GCATGGAGT→GCATGGAT
GCCTGGTGG→GCCTGGTG
GCTCGGGGT→GCTCGGGT
GGCAGGACA→GGCAGGAA
GGCTGGACA→GGCTGGAA
GGCTTGGTA→GGCTGGTA
GGTCGGTTA→GGTCGGTA
GGTTGGTTA→GGTTGGTA
GTCTGGTGA→GTCTGGTA
TAAAAGGTA→TAAAGGTA
TAACGGTAA→TAACGGTA
TAACTGGCA→TAACGGCA
TAATGGATA→TAATGGAA
TAATTGGGA→TAATGGGA
TAGCGGCGA→TAGCGGCA
TATAGGTGT→TATAGGTT
TATTGGTGT→TATTGGTT
TCCGGTGAG→TCC-GGTG
TGACGGTTG→TGACGGTG
TGCCGGTGG→TGCCGGTG
TGCTTGGTG→TGCTGGTG
TTAATGGCA→TTAAGGCA
AAAACGGATA→AAAAGGAA
AAACAGGTAA→AAACGGTA
AAATAGGTAT→AAATGGTT
AAATGGATGA→AAATGGAA
AAATTGGCAG→AAATGGCG
AACAAGGTTA→AACAGGTA
AACATGGCAG→AACAGGCG
AACCAGGTAG→AACCGGTG
AACCGGCAGC→AACCGGCC
AACCTGGCAA→AACCGGCA

AGTTTGTTA→AGTTGTTA
ATAAGGCAT→ATAAGGCT
ATAGGCCTA→ATA-GGCA
ATCAGGAAG→ATCAGGAG
ATCAGGTTA→ATCAGGTA
ATCCGGCAA→ATCCGGCA
ATCCGGTTA→ATCCGGTA
ATCGGGACA→ATCGGGAA
ATCGGTTTA→ATC-GGTA
ATCRGGATA→ATCRGGAA
ATCTGGCGG→ATCTGGCG
ATCTGGTGT→ATCTGGTT
ATGCGGACA→ATGCGGAA
ATGTGGTCA→ATGTGGTA
ATTATGGCA→ATTAGGCA
ATTCTGGTA→ATTCGGTA
ATTTGGATA→ATTTGGAA
ATTTTGGAA→ATTTGGAA
CAACGGCCA→CAACGGCA
CAACTGGTA→CAACGGTA
CAATCGGTG→CAATGGTG
CAATTGGCA→CAATGGCA
CACTGGTGG→CACTGGTG
CAGCGGTGA→CAGCGGTA
CAGCTGGTA→CAGCGGTA
CAGTGGCGA→CAGTGGCA
CATTGGTTA→CATTGGTA
CGAATGGTA→CGAAGGTA
CGCTGGTGC→CGCTGGTC
CGTATGGAT→CGTAGGAT
CTAATGGGA→CTAAGGGA
CTCTGGTGT→CTCTGGTT
GAACAGTAG→GAACAGTG
GAATGGCAT→GAATGGCT
GAATTGCAG→GAATGCAG
GACCGGATA→GACCGGAA
GAGAGGCGA→GAGAGGCA
GAGTGGCGA→GAGTGGCA
GATGGGCGA→GATGGGCA
GATTGGCAG→GATTGGCG
GCATGGGGT→GCATGGGT
GCGTTGGGC→GCGTGGGC
GCTGGCAGG→GCT-GGCG
GGCAGGATA→GGCAGGAA
GGCTGGATA→GGCTGGAA
GGGCGGTGC→GGGCGGTC
GGTCTGGTA→GGTCGGTA
GGTTGGTTG→GGTTGGTG
GTCTGGTGG→GTCTGGTG
TAAATGGTA→TAAAGGTA
TAACGGTAG→TAACGGTG
TAACTGGGA→TAACGGGA
TAATGGCTA→TAATGGCA
TAATTGGTA→TAATGGTA
TAGCGGTAA→TAGCGGTA
TATCGGTAG→TATCGGTG
TATTGGTTG→TATTGGTG
TCCNGGTGA→TCCNGGTA
TGAGTCTGA→TGA-GTCA
TGCCTGGTA→TGCCGGTA
TGTGGCCAA→TGT-GGCA
TTAATGGGA→TTAAGGGA
AAAACGGTGA→AAAAGGTA
AAACTGGTAG→AAACGGTG
AAATCGGTAG→AAATGGTG
AAATGGCCAG→AAATGGCG
AAATTGGTAA→AAATGGTA
AACACGGTTA→AACAGGTA
AACATGGTAG→AACAGGTG
AACCCGGCAG→AACCGGCG
AACCGGCAGG→AACCGGCG
AACCTGGCAG→AACCGGCG

AGTYGGTTA→AGTYGGTA
ATAAGGTAT→ATAAGGTT
ATAGGCTTA→ATA-GGCA
ATCAGGACA→ATCAGGAA
ATCATGGAA→ATCAGGAA
ATCCGGCCA→ATCCGGCA
ATCGGCACG→ATC-GGCG
ATCGGGATA→ATCGGGAA
ATCGTGACA→ATCGGACA
ATCTAGGGA→ATCTGGGA
ATCTGGCTA→ATCTGGCA
ATCTGGTTA→ATCTGGTA
ATGCGGATA→ATGCGGAA
ATGTGGTTA→ATGTGGTA
ATTCGGATA→ATTCGGAA
ATTGGCGCA→ATT-GGCA
ATTTGGCTA→ATTTGGCA
CAAAAGGCG→CAAAGGCG
CAACGGCTA→CAACGGCA
CAAGAGGTG→CAAGGGTG
CAATGGCTA→CAATGGCA
CAATTGGTA→CAATGGTA
CACTGGTGT→CACTGGTT
CAGCGGTTA→CAGCGGTA
CAGGGGATA→CAGGGGAA
CAGTTGGCA→CAGTGGCA
CCATGGTGG→CCATGGTG
CGCCCGGTA→CGCCGGTA
CGCTGGTGG→CGCTGGTG
CGTCTGGTA→CGTCGGTA
CTCCGGTGA→CTCCGGTA
CTGTAGGAG→CTGTGGAG
GAACAGTCA→GAACAGTA
GAATGGCTA→GAATGGCA
GAATTGGCA→GAATGGCA
GACCGGCTA→GACCGGCA
GAGCGGCAA→GAGCGGCA
GATAGGTGA→GATAGGTA
GATGTAGAA→GAT-GTAA
GATTGGGAG→GATTGGGG
GCCCGGTGG→GCCCGGTG
GCTAGGAGT→GCTAGGAT
GCTTGGTAA→GCTTGGTA
GGCAGGTTG→GGCAGGTG
GGCTGGTGA→GGCTGGTA
GGGTGGTCC→GGGTGGTC
GGTGGCTCA→GGT-GGCA
GGTTTGGTA→GGTTGGTA
GTCAAGGTGA→TCAAGGTA
TCCTGGTGG→TCCTGGTG
TGATGGTTG→TGATGGTG
TGCTGGATG→TGCTGGAG
TGTTGGAAG→TGTTGGAG
TTCCGGTGA→TTCCGGTA
AAAAGGATAA→AAAAGGAA
AAAGGGCTAT→AAAGGGCT
AAATCTGGCA→AAATGGCA
AAATGGCTAA→AAATGGCA
AAATTGGTAG→AAATGGTG
AACAGGAAAT→AACAGGAT
AACCAGGCAG→AACCGGCG
AACCCGGTAA→AACCGGTA
AACCGGCAGT→AACCGGCT
AACCTGGTAG→AACCGGTG

AGYTGGTTA→AGYTGGTA
ATAAGGTTA→ATAAGGTA
ATATGGTAC→ATATGGTC
ATCAGGAGA→ATCAGGAA
ATCCGGAAA→ATCCGGAA
ATCCGGCTA→ATCCGGCA
ATCGGCCTT→ATC-GGCT
ATCGGGTTA→ATCGGGTA
ATCGTGATA→ATCGGATA
ATCTGGAAA→ATCTGGAA
ATCTGGTCA→ATCTGGTA
ATCTGGTTG→ATCTGGTG
ATGCGGTAG→ATGCGGTG
ATGTTGGAA→ATGTGGAA
ATTCGGGTA→ATTCGGGA
ATTGGTTCC→ATT-GGTC
ATTTGGTAA→ATTTGGTA
CAACAGGTA→CAACGGTA
CAACGGTCA→CAACGGTA
CAAGCGGTG→CAAGGGTG
CAATGGTCA→CAATGGTA
CAATTGTCA→CAATGTCA
CAGCAGGTA→CAGCGGTA
CAGCTAGGA→CAGCGG-A
CAGGTGGTA→CAGGGGTA
CAGTTGGGA→CAGTGGGA
CCCCGGTGG→CCCCGGTG
CGCGGCCTA→CGC-GGCA
CGGAATTGG→CGGAATTG
CGTTGGCTA→CGTTGGCA
CTCCGGTGG→CTCCGGTG
CTGTTGGCG→CTGTGGCG
GAACGGCTA→GAACGGCA
GAATGGTAT→GAATGGTT
GAATTGGTA→GAATGGTA
GAGTGGGA→GAGCGGGA
GATCGAGAA→GATCGAGA
GATTCAGAA→GATTCAGA
GATTGGTAG→GATTGGTG
GCCGGAATA→GCC-GGAA
GCTAGGGAG→GCTAGGGG
GCTTGGTGC→GCTTGGTC
GGCGGCCTA→GGC-GGCA
GGCTGGTGC→GGCTGGTC
GGGTGGTTA→GGGTGGTA
GGTGGCTGA→GGT-GGCA
GTAGAGGTA→GTAGGGTA
GTTACGGAG→GTTAGGAG
TAACGGCCA→TAACGGCA
TAACGGTCA→TAACGGTA
TAATAGGGA→TAATGGGA
TAATTGGAA→TAATGGAA
TACTGGTGT→TACTGGTT
TAGTGGTAA→TAGTGGTA
TATTGGTAG→TATTGGTG
TCAGCGGGA→TCAGGGGA
TCTGGTAGA→TCT-GGTA
TGCAGGTAG→TGCAGGTG
TGCTGGTGC→TGCTGGTC
TGTTGGTTA→TGTTGGTA
TTCTGGTGA→TTCTGGTA
AAAAGGTTTA→AAAAGGTA
AAATAGGATA→AAATGGAA
AAATGGAATA→AAATGGAA
AAATTGGAAA→AAATGGAA
AACAAGGCAA→AACAGGCA
AACAGGCAGT→AACAGGCT
AACCAGGCTA→AACCGGCA
AACCCGGTAT→AACCGGTT
AACCGGTTAG→AACCGGTG
AACCTGGTAT→AACCGGTT

AKTTGGTTA→AKTTGGTA
ATACGGATA→ATACGGAA
ATATGGTTA→ATATGGTA
ATCAGGATA→ATCAGGAA
ATCCGGACA→ATCCGGAA
ATCCGGTCA→ATCCGGTA
ATCGGCGAA→ATC-GGCA
ATCGGTTAA→ATC-GGTA
ATCNGGATA→ATCNGGAA
ATCTGGATA→ATCTGGAA
ATCTGGTGA→ATCTGGTA
ATCTTGGTA→ATCTGGTA
ATGTCGGAA→ATGTGGAA
ATGTTGGCT→ATGTGGCT
ATTCGGTAT→ATTCGGTT
ATTGGTTTC→ATT-GGTC
ATTTGGTAT→ATTTGGTT
CAACAGTCA→CAACGTCA
CAACGGTTA→CAACGGTA
CAATAGGTA→CAATGGTA
CAATGGTTA→CAATGGTA
CACCGGCGA→CACCGGCA
CAGCGGCAG→CAGCGGCG
CAGCTGGAA→CAGCGGAA
CAGGTTGTA→CAGGTGTA
CAGTTGGTA→CAGTGGTA
CCCTGGTGG→CCCTGGTG
CGCGGGATA→CGCGGGAA
CGGGGGTTC→CGGGGGTC
CGTTGGTTA→CGTTGGTA
CTCTGGTGA→CTCTGGTA
GAAAGGCAA→GAAAGGCA
GAACTGGTA→GAACGGTA
GAATGGTGA→GAATGGTA
GACAGGCCC→GACAGGCC
GACTGGTGC→GACTGGTC
GAGGGGTTA→GAGGGGTA
GATGCAGAA→GAT-GCAA
GATTGAGAA→GATTGAGA
GATTGGTGA→GATTGGTA
GCCGGATTC→GCC-GGAC
GCTAGGGGT→GCTAGGGT
GGACGGTCA→GGACGGTA
GGCTGGAAG→GGCTGGAG
GGCTGGTTA→GGCTGGTA
GGGTTGGGA→GGGTGGGA
GGTTGGTCA→GGTTGGTA
GTAGGGGTA→GTAGGGGA
NGCTGGATA→NGCTGGAA
TAACGGCTA→TAACGGCA
TAACGGTTA→TAACGGTA
TAATCGGCA→TAATGGCA
TAATTGGCA→TAATGGCA
TAGCGGAAA→TAGCGGAA
TAGTTGGTA→TAGTGGTA
TATTGGTGC→TATTGGTC
TCCCGGTGG→TCCCGGTG
TCTGGTTGA→TCT-GGTA
TGCCGGTGC→TGCCGGTC
TGCTGGTGT→TGCTGGTT
TTAACGGCA→TTAAGGCA
TTCTGGTGT→TTCTGGTT
AAAATGGATA→AAAAGGAA
AAATAGGTAG→AAATGGTG
AAATGGAGAT→AAATGAGT
AAATTGGAAG→AAATGGAG
AACAAGGTAG→AACAGGTG
AACAGGTTCA→AACAGGTA
AACCAGGTAA→AACCGGTA
AACCCGGTTA→AACCGGTA
AACCTGGATA→AACCGGAA
AACCTGGTCA→AACCGGTA

AACCTGGTTA→AACCGGTA
AACGGCTTCA→AAC-GGCA
AACGGTTTTA→AAC-GGTA
AACTAGGCTA→AACTGGCA
AACTCTGGCA→AACTGGCA
AACTGGATCA→AACTGGAA
AACTGGCAGG→AACTGGCG
AACTGGCTTT→AACTGGCT
AACTGGTGAG→AACTGGTG
AACTGGTTTA→AACTGGTA
AACTTGGCAG→AACTGGCG
AACTTTGGAA→AACTGGAA
AAGCTGGATA→AAGCGGAA
AAGTTGGTTA→AAGTGGTA
AATACGGTAG→AATAGGTG
AATAGGTCAG→AATAGGTG
AATATGGTAT→AATAGGTT
AATCAGGTAG→AATCGGTG
AATCGGCAGC→AATCGGCC
AATCTGGCAA→AATCGGCA
AATCTGGTTA→AATCGGTA
AATGGCATCC→AAT-GGCC
AATTAGGTAT→AATTGGTT
AATTGGAAAG→AATTGGAG
AATTGGATCA→AATTGGAA
AATTGGCAGC→AATTGGCC
AATTGGCCGA→AATTGGCA
AATTGGCTGA→AATTGGCA
AATTGGGTAG→AATTGGGG
AATTGGTAGG→AATTGGTG
AATTGGTGTA→AATTGGTA
AATTGGTTTT→AATTGGTT
AATTTGGTAA→AATTGGTA
AATTTGGTTA→AATTGGTA
ACGTTGGCAA→ACGTGGCA
AGAAGGTCGA→AGAAGGTA
AGACGGCCGA→AGACGGCA
AGACTGGTCA→AGACGGTA
AGAGGCCCAG→AGA-GGCG
AGAGGGCCTA→AGA-GGGA
AGAGGTTATA→AGA-GGTA
AGATGGACCA→AGATGGAA
AGATGGATTA→AGATGGAA
AGATGGGATA→AGATGGGA
AGATTGGCCA→AGATGGCA
AGCAAGGTTA→AGCAGGTA
AGCAGGAGAG→AGCAGGAG
AGCAGGCTGA→AGCAGGCA
AGCATGGAGT→AGCAGGAT
AGCATGGTTT→AGCAGGTT
AGCCAGGAGT→AGCCGGAT
AGCCAGGGTA→AGCCGGGA
AGCCAGGTGA→AGCCGGTA
AGCCCGGATA→AGCCGGAA
AGCCCGGTCG→AGCCGGTG
AGCCCGGCCA→AGCCGGCA
AGCCGGTCGA→AGCCGGTA
AGCCTGGAAT→AGCCGGAT
AGCCTGGCAA→AGCCGGCA
AGCCTGGGTA→AGCCGGGA
AGCCTGGTGA→AGCCGGTA
AGCGCGGTCA→AGCGGGTA
AGCGGCATTA→AGC-GGCA
AGCGGCTCAA→AGC-GGCA
AGCGGTCACA→AGC-GGTA
AGCGGTTTTA→AGC-GGTA
AGCTAGGAGT→AGCTGGAT
AGCTAGGTAA→AGCTGGTA
AGCTCGGCCA→AGCTGGCA
AGCTGGACAA→AGCTGGAA
AGCTGGATAG→AGCTGGAG
AGCTGGCAAG→AGCTGGCG

AACGAGTCCA→AACGGTCA
AACGGTATTA→AAC-GGTA
AACGTGGTAG→AACGGGTG
AACTAGGTAG→AACTGGTG
AACTGGAAGA→AACTGGAA
AACTGGATGA→AACTGGAA
AACTGGCAGT→AACTGGCT
AACTGGGCAG→AACTGGGG
AACTGGTGTT→AACTGGTT
AACTGGTTTG→AACTGGTG
AACTTGGCAT→AACTGGCT
AAGAAGGTCA→AAGAGGTA
AAGGGAATTA→AAG-GGAA
AATAAGGCAG→AATAGGCG
AATAGGCAGC→AATAGGCC
AATAGGTGAT→AATAGGTT
AATCAGGAAG→AATCGGAG
AATCAGGTAT→AATCGGTT
AATCGGCAGG→AATCGGCG
AATCTGGCAG→AATCGGCG
AATGAGGTTA→AATGGGTA
AATTAGGATA→AATTGGAA
AATTCGGCAA→AATTGGCA
AATTGGAAGA→AATTGGAA
AATTGGATTA→AATTGGAA
AATTGGCAGG→AATTGGCG
AATTGGCGAT→AATTGGCT
AATTGGGAAG→AATTGGGG
AATTGGTAAT→AATTGGTT
AATTGGTAGT→AATTGGTT
AATTGGTTAA→AATTGGTA
AATTGTGGCA→AATTGGCA
AATTTGGTAC→AATTGGTC
ACAAGGTGTT→ACAAGGTT
ACTGGGTTTA→ACTGGGTA
AGAATGGACA→AGAAGGAA
AGACGGCCTA→AGACGGCA
AGACTGGTTA→AGACGGTA
AGAGGCCCCA→AGA-GGCA
AGAGGTCATA→AGA-GGTA
AGAGGTTGAC→AGA-GGTC
AGATGGAGAA→AGATGGAA
AGATGGCCGA→AGATGGCA
AGATGGGTTA→AGATGGTA
AGATTGGCTA→AGATGGCA
AGCACGGCAC→AGCAGGCC
AGCAGGATTA→AGCAGGAA
AGCAGGTCAA→AGCAGGTA
AGCATGGATA→AGCAGGAA
AGCCAGGAAA→AGCCGGAA
AGCCAGGATA→AGCCGGAA
AGCCAGGTAA→AGCCGGTA
AGCCAGGTTA→AGCCGGTA
AGCCCGGCAA→AGCCGGCA
AGCCCGGTGA→AGCCGGTA
AGCCGGCCGA→AGCCGGCA
AGCCGGTTCA→AGCCGGTA
AGCCTGGACA→AGCCGGAA
AGCCTGGCCA→AGCCGGCA
AGCCTGGTAA→AGCCGGTA
AGCCTGGTTA→AGCCGGTA
AGCGGAATAA→AGC-GGAA
AGCGGCCAAA→AGC-GGCA
AGCGGGAGAG→AGCGGGAG
AGCGGTCATA→AGC-GGTA
AGCGTGGATA→AGCGGGAA
AGCTAGGATA→AGCTGGAA
AGCTAGGTAT→AGCTGGTT
AGCTCGGTCA→AGCTGGTA
AGCTGGACCA→AGCTGGAA
AGCTGGATAN→AGCTGGAN
AGCTGGCCAA→AGCTGGCA

AACGGAACCA→AAC-GGAA
AACGGTTAGG→AAC-GGTG
AACGTTGGCA→AACGGGCA
AACTAGGTTA→AACTGGTA
AACTGGAAGG→AACTGGAG
AACTGGCAAG→AACTGGCG
AACTGGCCAG→AACTGGCG
AACTGGTAGG→AACTGGTG
AACTGGTTAG→AACTGGTG
AACTGGTTTT→AACTGGTT
AACTTGGTAG→AACTGGTG
AAGCCGGATA→AAGCGGAA
AAGTGGCTAA→AAGTGGCA
AATAAGGCTA→AATAGGCA
AATAGGCAGT→AATAGGCT
AATAGGTTTG→AATAGGTG
AATCAGGATA→AATCGGAA
AATCCGGAAG→AATCGGAG
AATCGGTAGC→AATCGGTC
AATCGGTCAG→AATCGGTG
AATGATGGTA→AATGGGTA
AATTAGGCAG→AATTGGCG
AATTCGGGCA→AATTGGGA
AATTGGAATA→AATTGGAA
AATTGGCAAG→AATTGGCG
AATTGGCAGT→AATTGGCT
AATTGGCTAA→AATTGGCA
AATTGGGATA→AATTGGAA
AATTGGTACG→AATTGGTG
AATTGGTCAG→AATTGGTG
AATTGGTTAG→AATTGGTG
AATTTGGATA→AATTGGAA
AATTTGGTAG→AATTGGTG
ACCGGATGTT→ACC-GGAT
ACTGGTTGTG→ACT-GGTG
AGACAGGCTA→AGACGGCA
AGACGGCGAG→AGACGGCG
AGAGGAAATG→AGA-GGAG
AGAGGCCGTA→AGA-GGCA
AGAGGTCGAC→AGA-GGTC
AGATAGGTTA→AGATGGTA
AGATGGAGAG→AGATGGAG
AGATGGCCTA→AGATGGCA
AGATGGTCGA→AGATGGTA
AGATTGGTCA→AGATGGTA
AGCACGGCCA→AGCAGGCA
AGCAGGCCAG→AGCAGGCG
AGCAGGTCGA→AGCAGGTA
AGCATGGCTA→AGCAGGCA
AGCCAGGACA→AGCCGGAA
AGCCAGGATG→AGCCGGAG
AGCCAGGTAG→AGCCGGTG
AGCCATGGTA→AGCCGGTA
AGCCCGGCCA→AGCCGGCA
AGCCCGGTTA→AGCCGGTA
AGCCGGCCTA→AGCCGGCA
AGCCGGTTTA→AGCCGGTA
AGCCTGGAGT→AGCCGGAT
AGCCTGGCTA→AGCCGGCA
AGCCTGGTAG→AGCCGGTG
AGCCTGGTTG→AGCCGGTG
AGCGGACTTA→AGC-GGAA
AGCGGCCATA→AGC-GGCA
AGCGGGAGCG→AGCGGGAG
AGCGGTCGAC→AGC-GGTC
AGCGTGGTCA→AGCGGGTA
AGCTAGGATG→AGCTGGAG
AGCTAGGTTA→AGCTGGTA
AGCTCGGTTA→AGCTGGTA
AGCTGGACGA→AGCTGGAA
AGCTGGATAT→AGCTGGAT
AGCTGGCCAG→AGCTGGCG

AACGGAATTA→AAC-GGAA
AACGGTTATA→AAC-GGTA
AACTAGGCAA→AACTGGCA
AACTCGGAGA→AACTGGAA
AACTGGATAA→AACTGGAA
AACTGGCACG→AACTGGCG
AACTGGCGAG→AACTGGCG
AACTGGTCAG→AACTGGTG
AACTGGTTAT→AACTGGTT
AACTTGGATA→AACTGGAA
AACTTGGTAT→AACTGGTT
AAGCCGGTTA→AAGCGGTA
AAGTGGTTTG→AAGTGGTG
AATAAGGTAT→AATAGGTT
AATAGGCGAG→AATAGGCG
AATATGGATA→AATAGGAA
AATCAGGCAG→AATCGGCG
AATCGGTAG→AATCGGTG
AATCGGTTAG→AATCGGTG
AATCTGGTAT→AATCGGTT
AATGGACCCA→AAT-GGAA
AATTAGGTAA→AATTGGTA
AATTCGGTAG→AATTGGTG
AATTGGACTA→AATTGGAA
AATTGGCAAT→AATTGGCT
AATTGGCATG→AATTGGCG
AATTGGCTAG→AATTGGCG
AATTGGGCAG→AATTGGGG
AATTGGTAGA→AATTGGTA
AATTGGTCAT→AATTGGTT
AATTGGTTAT→AATTGGTT
AATTTGGCAG→AATTGGCG
AATTTGGTAT→AATTGGTT
ACGAGGCCCA→ACGAGGCA
ACTTGGTCCG→ACTTGGTG
AGACAGGTCA→AGACGGTA
AGACGGCTTA→AGACGGCA
AGAGGACCTA→AGA-GGAA
AGAGGCTATA→AGA-GGCA
AGAGGTCGTA→AGA-GGTA
AGATCGGCAC→AGATGGCC
AGATGGATAG→AGATGGAG
AGATGGCTGA→AGATGGCA
AGATGGTGAG→AGATGGTG
AGATTGGTGA→AGATGGTA
AGCACGGTCA→AGCAGGTA
AGCAGGCCGA→AGCAGGCA
AGCAGGTCTA→AGCAGGTA
AGCATGGTCA→AGCAGGTA
AGCCAGGACG→AGCCGGAG
AGCCAGGCCA→AGCCGGCA
AGCCAGGTCA→AGCCGGTA
AGCCCGGAAA→AGCCGGAA
AGCCCGGCTA→AGCCGGCA
AGCCCGGTTC→AGCCGGTC
AGCCGGCTGA→AGCCGGCA
AGCCTGATTA→AGCCGATA
AGCCTGGATA→AGCCGGAA
AGCCTGGCTG→AGCCGGCG
AGCCTGGTAT→AGCCGGTT
AGCCTTGGTA→AGCCGGTA
AGCGGATTCA→AGC-GGAA
AGCGGCCGTA→AGC-GGCA
AGCGGGCCAA→AGCGGGCA
AGCGGTCGTA→AGC-GGTA
AGCGTGGTTA→AGCGGGTA
AGCTAGGCTA→AGCTGGCA
AGCTCGGACA→AGCTGGAA
AGCTGAGATA→AGCTGAGA
AGCTGGACTA→AGCTGGAA
AGCTGGATCA→AGCTGGAA
AGCTGGCCCA→AGCTGGCA

AACGGATTTA→AAC-GGAA
AACGGTTTCA→AAC-GGTA
AACTAGGCAT→AACTGGCT
AACTCGGTTA→AACTGGTA
AACTGGATAG→AACTGGAG
AACTGGCAGC→AACTGGCC
AACTGGCGCT→AACTGGCT
AACTGGTCAT→AACTGGTT
AACTGGTTGA→AACTGGTA
AACTTGGCAA→AACTGGCA
AACTTGGTTA→AACTGGTA
AAGCGGTCAA→AAGCGGTA
AAGTTGGCTA→AAGTGGCA
AATACGGATA→AATAGGAA
AATAGGCGAT→AATAGGCT
AATATGGTAG→AATAGGTG
AATCAGGTAA→AATCGGTA
AATCCGGTTA→AATCGGTA
AATCTGGATA→AATCGGAA
AATCTGGTCA→AATCGGTA
AATGGACTCA→AAT-GGAA
AATTAGGTAG→AATTGGTG
AATTCGGTTA→AATTGGTA
AATTGGATAG→AATTGGAG
AATTGGCACG→AATTGGCG
AATTGGCCAG→AATTGGCG
AATTGGCTAT→AATTGGCT
AATTGGGCAT→AATTGGGT
AATTGGTAGC→AATTGGTC
AATTGGTGAG→AATTGGTG
AATTGGTTTA→AATTGGTA
AATTTGGCTA→AATTGGCA
AATTTGGTCA→AATTGGTA
ACGCCCGGTA→ACGCGGTA
AGAAGGCCGA→AGAAGGCA
AGACCGGTTA→AGACGGTA
AGACTGGATA→AGACGGAA
AGAGGCCATA→AGA-GGCA
AGAGGCTGTA→AGA-GGCA
AGAGGTCGTC→AGA-GGTC
AGATCGGTTA→AGATGGTA
AGATGGATCA→AGATGGAA
AGATGGCTTA→AGATGGCA
AGATGGTTTA→AGATGGTA
AGATTGGTTA→AGATGGTA
AGCACGGTTA→AGCAGGTA
AGCAGGCTAG→AGCAGGCG
AGCAGGTTTA→AGCAGGTA
AGCATGGTTA→AGCAGGTA
AGCCAGGAGA→AGCCGGAA
AGCCAGGCTA→AGCCGGCA
AGCCAGGTCG→AGCCGGTG
AGCCCGGACA→AGCCGGAA
AGCCCGGTCA→AGCCGGTA
AGCCGGCCAA→AGCCGGCA
AGCCGGGTTA→AGCCGGGA
AGCCTGGAAA→AGCCGGAA
AGCCTGGATG→AGCCGGAG
AGCCTGGGAG→AGCCGGGG
AGCCTGGTCA→AGCCGGTA
AGCGAGTTCA→AGCGGTTA
AGCGGATTTA→AGC-GGAA
AGCGGCTATA→AGC-GGCA
AGCGGTAAAG→AGC-GGTG
AGCGGTTATA→AGC-GGTA
AGCTAGGAAA→AGCTGGAA
AGCTAGGGCT→AGCTGGGT
AGCTCGGAGT→AGCTGGAT
AGCTGGAATA→AGCTGGAA
AGCTGGAGAG→AGCTGGAG
AGCTGGATTA→AGCTGGAA
AGCTGGCCGA→AGCTGGCA

Chapter V – Appendix

AGCTGGCCTA→AGCTGGCA AGCTGGCGAG→AGCTGGCG AGCTGGCTAG→AGCTGGCG AGCTGGCTCA→AGCTGGCA AGCTGGCTGA→AGCTGGCA
AGCTGGCTGG→AGCTGGCG AGCTGGCTTA→AGCTGGCA AGCTGGGTGA→AGCTGGGA AGCTGGGTTA→AGCTGGTA AGCTGGTAGA→AGCTGGTA
AGCTGGTATA→AGCTGGTA AGCTGGTCAA→AGCTGGTA AGCTGGTCCA→AGCTGGTA AGCTGGTCGA→AGCTGGTA AGCTGGTCTA→AGCTGGTA
AGCTGGTTAG→AGCTGGTG AGCTGGTTCA→AGCTGGTA AGCTGGTTGA→AGCTGGTA AGCTGGTTTA→AGCTGGTA AGCTGTCCGA→AGCTGTCA
AGCTTGATTA→AGCTGATA AGCTTGGACA→AGCTGGAA AGCTTGGAGT→AGCTGGAT AGCTTGGATA→AGCTGGAA AGCTTGGCCA→AGCTGGCA
AGCTTGGCGA→AGCTGGCA AGCTTGGCTA→AGCTGGCA AGCTTGGCTG→AGCTGGCG AGCTTGGGCA→AGCTGGGA AGCTTGGGTA→AGCTGGGA
AGCTTGGTAA→AGCTGGTA AGCTTGGTCA→AGCTGGTA AGCTTGGTCC→AGCTGGTC AGCTTGGTTA→AGCTGGTA AGCTTGGTTG→AGCTGGTG
AGGCAGGTCA→AGGCGGTA AGGCAGGTTA→AGGCGGTA AGGCCGGTCG→AGGCGGTG AGGCGGACGA→AGGCGGAA AGGCGGTTTA→AGGCGGTA
AGGCTGGATA→AGGCGGAA AGGCTGGCCA→AGGCGGCA AGGCTGGCTA→AGGCGGCA AGGCTGGTTA→AGGCGGTA AGGGGGCTGA→AGGGGGCA
AGGGGGGGTA→AGGGGGGA AGGGGTCATA→AGG-GGTA AGGTAGGATA→AGGTGGAA AGGTCGGAGT→AGGTGGAT AGGTCGGATA→AGGTGGAA
AGGTCGGTCA→AGGTGGTA AGGTCGGTTA→AGGTGGTA AGGTGGAGAG→AGGTGGAG AGGTGGCGAG→AGGTGGCG AGGTGGCTTA→AGGTGGCA
AGGTGGTCGA→AGGTGGTA AGGTGGTTAG→AGGTGGTG AGGTGGTTTA→AGGTGGTA AGGTTGGCTA→AGGTGGCA AGGTTGGTCA→AGGTGGTA
AGGTTGGTTA→AGGTGGTA AGTAAGGTTA→AGTAGGTA AGTACGGTAA→AGTAGGTA AGTACGGTTA→AGTAGGTA AGTAGGAGAG→AGTAGGAG
AGTAGGATCA→AGTAGGAA AGTAGGCCAA→AGTAGGCA AGTAGGCCGA→AGTAGGCA AGTAGGCTCA→AGTAGGCA AGTAGGCTGA→AGTAGGCA
AGTAGGCTTA→AGTAGGCA AGTAGGTCGA→AGTAGGTA AGTAGGTTCA→AGTAGGTA AGTAGGTTGA→AGTAGGTA AGTAGGTTTA→AGTAGGTA
AGTATGGAGT→AGTAGGAT AGTATGGATA→AGTAGGAA AGTATGGTTA→AGTAGGTA AGTCAGGAAA→AGTCGGAA AGTCAGGACA→AGTCGGAA
AGTCAGGAGA→AGTCGGAA AGTCAGGAGT→AGTCGGAT AGTCAGGATA→AGTCGGAA AGTCAGGCAA→AGTCGGCA AGTCAGGCCA→AGTCGGCA
AGTCAGGCTA→AGTCGGCA AGTCAGGGGA→AGTCGGGA AGTCAGGTAA→AGTCGGTA AGTCAGGTCA→AGTCGGTA AGTCAGGTGA→AGTCGGTA
AGTCAGGTTA→AGTCGGTA AGTCCGGAAT→AGTCGGAT AGTCCGGACA→AGTCGGAA AGTCCGGAGT→AGTCGGAT AGTCCGGATA→AGTCGGAA
AGTCCGGCCA→AGTCGGCA AGTCCGGCCG→AGTCGGCG AGTCCGGCTA→AGTCGGCA AGTCCGGTCA→AGTCGGTA AGTCCGGTGA→AGTCGGTA
AGTCCGGTTA→AGTCGGTA AGTCGGAGCA→AGTCGGAA AGTCGGCCAA→AGTCGGCA AGTCGGCCCA→AGTCGGCA AGTCGGCCGA→AGTCGGCA

AGTCGGCCTA→AGTCGGCA AGTCGGCTGA→AGTCGGCA AGTCGGCTTA→AGTCGGCA AGTCGGTCAA→AGTCGGTA AGTCGGTCGA→AGTCGGTA
AGTCGGTCTA→AGTCGGTA AGTCGGTTCA→AGTCGGTA AGTCGGTTGA→AGTCGGTA AGTCGGTTTA→AGTCGGTA AGTCTGATCA→AGTCGATA
AGTCTGATTA→AGTCGATA AGTCTGCGCA→AGTCGCGA AGTCTGGAAA→AGTCGGAA AGTCTGGACA→AGTCGGAA AGTCTGGACG→AGTCGGAG
AGTCTGGAGA→AGTCGGAA AGTCTGGAGT→AGTCGGAT AGTCTGGATA→AGTCGGAA AGTCTGGCCA→AGTCGGCA AGTCTGGCTA→AGTCGGCA
AGTCTGGGCA→AGTCGGGA AGTCTGGTAA→AGTCGGTA AGTCTGGTAG→AGTCGGTG AGTCTGGTCA→AGTCGGTA AGTCTGGTCG→AGTCGGTG
AGTCTGGTGA→AGTCGGTA AGTCTGGTTA→AGTCGGTA AGTCTGGTTC→AGTCGGTC AGTGCGGTTA→AGTGGGTA AGTGGAACCA→AGT-GGAA
AGTGGAATCA→AGT-GGAA AGTGGACTCA→AGT-GGAA AGTGGAGTGA→AGT-GGAA AGTGGATAAA→AGT-GGAA AGTGGATATA→AGT-GGAA
AGTGGATGCA→AGT-GGAA AGTGGATTCA→AGT-GGAA AGTGGATTTA→AGT-GGAA AGTGGCCAAA→AGT-GGCA AGTGGCCATA→AGT-GGCA
AGTGGCCCTA→AGT-GGCA AGTGGCCGTA→AGT-GGCA AGTGGCCTAA→AGT-GGCA AGTGGCTATA→AGT-GGCA AGTGGCTGTA→AGT-GGCA
AGTGGCTTAA→AGT-GGCA AGTGGGAGAG→AGTGGGAG AGTGGGCCAA→AGTGGGCA AGTGGGCCCA→AGTGGGCA AGTGGGCTGA→AGTGGGCA
AGTGGTCACA→AGT-GGTA AGTGGTCATA→AGT-GGTA AGTGGTCGAC→AGT-GGTC AGTGGTCGCA→AGT-GGTA AGTGGTCGTA→AGT-GGTA
AGTGGTTAAA→AGT-GGTA AGTGGTTAAC→AGT-GGTC AGTGGTTATA→AGT-GGTA AGTGGTTGAA→AGT-GGTA AGTGGTTGTA→AGT-GGTA
AGTGGTTGTC→AGT-GGTC AGTGGTTTCT→AGT-GGTT AGTTAGCGTA→AGTTGCGA AGTTAGGAAA→AGTTGGAA AGTTAGGACA→AGTTGGAA
AGTTAGGAGT→AGTTGGAT AGTTAGGATA→AGTTGGAA AGTTAGGTAA→AGTTGGTA AGTTAGGTTA→AGTTGGTA AGTTATGGTA→AGTTGGTA
AGTTCCGGAA→AGTTGGAA AGTTCGGACA→AGTTGGAA AGTTCGGAGT→AGTTGGAT AGTTCGGCTA→AGTTGGCA AGTTCGGTCA→AGTTGGTA
AGTTCGGTGA→AGTTGGTA AGTTCGGTTA→AGTTGGTA AGTTGAGAGT→AGTTGAGT AGTTGGAATA→AGTTGGAA AGTTGGACAA→AGTTGGAA
AGTTGGACGA→AGTTGGAA AGTTGGAGAG→AGTTGGAG AGTTGGAGCA→AGTTGGAA AGTTGGAGTA→AGTTGGAA AGTTGGATAA→AGTTGGAA
AGTTGGATCA→AGTTGGAA AGTTGGATGA→AGTTGGAA AGTTGGATTA→AGTTGGAA AGTTGGCATA→AGTTGGCA AGTTGGCCAA→AGTTGGCA
AGTTGGCCAG→AGTTGGCG AGTTGGCCCA→AGTTGGCA AGTTGGCCGA→AGTTGGCA AGTTGGCCTA→AGTTGGCA AGTTGGCGAG→AGTTGGCG
AGTTGGCTAA→AGTTGGCA AGTTGGCTGA→AGTTGGCA AGTTGGCTTA→AGTTGGCA AGTTGGGATA→AGTTGGAA AGTTGGGTAG→AGTTGGTG
AGTTGGNTTA→AGTTGGNA AGTTGGTAGA→AGTTGGTA AGTTGGTAGT→AGTTGGTT AGTTGGTCAA→AGTTGGTA AGTTGGTCCA→AGTTGGTA
AGTTGGTCGA→AGTTGGTA AGTTGGTCTA→AGTTGGTA AGTTGGTGGG→AGTTGGTG AGTTGGTGTA→AGTTGGTA AGTTGGTTAA→AGTTGGTA
AGTTGGTTAG→AGTTGGTG AGTTGGTTCA→AGTTGGTA AGTTGGTTGA→AGTTGGTA AGTTGGTTTA→AGTTGGTA AGTTTGATCA→AGTTGATA
AGTTTGATTA→AGTTGATA AGTTTGCGTA→AGTTGCGA AGTTTGGAAG→AGTTGGAG AGTTTGGACA→AGTTGGAA AGTTTGGAGT→AGTTGGAT
AGTTTGGATA→AGTTGGAA AGTTTGGCCA→AGTTGGCA AGTTTGGCGA→AGTTGGCA AGTTTGGCTA→AGTTGGCA AGTTTGGGTA→AGTTGGGA
AGTTTGGTAA→AGTTGGTA AGTTTGGTAC→AGTTGGTC AGTTTGGTAG→AGTTGGTG AGTTTGGTAT→AGTTGGTT AGTTTGGTCA→AGTTGGTA
AGTTTGGTTA→AGTTGGTA AGTTTGGTTG→AGTTGGTG AGTTTTGGTA→AGTTGGTA AGTTTTGTTA→AGTTGTTA ATAAGGATTG→ATAAGGAG
ATAATTGGTA→ATAAGGTA ATACGGCTGG→ATACGGCG ATACGGTTTG→ATACGGTG ATACTGACAT→ATACGACT ATATGGTTGG→ATATGGTG
ATATTGGTTA→ATATGGTA ATCAGGAATA→ATCAGGAA ATCTGGATCA→ATCTGGAA ATGCTGGTTA→ATGCGGTA ATGGCGGTTA→ATGGGGTA
ATGGTGGTTA→ATGGGGTA ATGTTGGCTT→ATGTGGCT ATGTTGGTTA→ATGTGGTA ATTATGGCAA→ATTAGGCA ATTGGTCTGA→ATT-GGTA
ATTTCCTGGT→ATTTGG-T ATTTGGTGTA→ATTTGGTA ATTTTGGATA→ATTTGGAA CAAAAGGCTA→CAAAGGCA CAAATGGTTA→CAAAGGTA
CAACAGGTTA→CAACGGTA CAACGGTTCA→CAACGGTA CAACTGGATA→CAACGGAA CAACTGGCTA→CAACGGCA CAACTGGCTG→CAACGGCG
CAAGTTCGGC→CAAGGG-C CAATGGTTAA→CAATGGTA CAATTGGATA→CAATGGAA CAATTGGCAG→CAATGGCG CAATTGGTTC→CAATGGTC
CAGAGGCAGG→CAGAGGCG CAGAGGTCCA→CAGAGGTA CAGAGGTCGT→CAGAGGTT CAGCAGGTCA→CAGCGGTA CAGCCTGGTA→CAGCGGTA
CAGCGGCCTA→CAGCGGCA CAGCGGCGAG→CAGCGGCG CAGCGGTCAA→CAGCGGTA CAGCTGGATA→CAGCGGAA CAGCTGGCTA→CAGCGGCA
CAGCTTGGTA→CAGCGGTA CAGTCGGTTA→CAGTGGTA CAGTTGCGTA→CAGTGCGA CAGTTGGATA→CAGTGGAA CAGTTGGTTA→CAGTGGTA
CAGTTTGGTA→CAGTGGTA CATCCGGTGA→CATCGGTA CCGAGTGGCC→CCGAGGCC CGACGGCGAG→CGACGGCG CGACGGCGTG→CGACGGCG
CGATGGAGAG→CGATGGAG CGATGGAGCG→CGATGGAG CGATGGCGAG→CGATGGCG CGATGGCGGG→CGATGGCG CGATGGTGAG→CGATGGTG
CGCCGGAGAG→CGCCGGAG CGCGGGAGAG→CGCGGGAG CGCTGGAGAG→CGCTGGAG CGGGACGGTA→CGGGGGTA CGTAGAGGTA→CGTAGGTA
CGTCGGAGAG→CGTCGGAG CGTGGGAGAG→CGTGGGAG CGTTGGAGAG→CGTTGGAG CGTTGGAGGG→CGTTGGAG CTAACGGCCA→CTAAGGCA
CTGAAGGCAG→CTGAGGCG CTGATGGCAG→CTGAGGCG CTGCTGGTAG→CTGCGGTG CTGCTGGTGG→CTGCGGTG CTGTTGGCAG→CTGTGGCG
CTGTTGGTAG→CTGTGGTG CTGTTGGTCG→CTGTGGTG GAAATGGTAG→GAAAGGTG GAACAGGTAT→GAACGGTT GAACTGGACT→GAACGGAT
GAACTGGCAG→GAACGGCG GAACTGGCTA→GAACGGCA GAACTGGTTA→GAACGGTA GAATCAGCAG→GAATAGCG GAATCGGCTA→GAATGGCA
GAATCGGTAG→GAATGGTG GAATGGCTTT→GAATGGCT GAATTGGCAG→GAATGGCG GAATTGGCTA→GAATGGCA GAATTGGTAG→GAATGGTG
GAATTGGTCA→GAATGGTA GAATTGGTTA→GAATGGTA GACTTGGGGA→GACTGGGA GAGAGGATGA→GAGAGGAA GAGAGGATTA→GAGAGGAA
GAGAGGCTGA→GAGAGGCA GAGAGGTTTA→GAGAGGTA GAGCGGCCAA→GAGCGGCA GAGCGGCTTA→GAGCGGCA GAGCGGTCAA→GAGCGGTA
GAGCGGTCGA→GAGCGGTA GAGCGGTCTA→GAGCGGTA GAGCGGTTTA→GAGCGGTA GAGTGGCCAA→GAGTGGCA GAGTGGCCGA→GAGTGGCA
GAGTGGCTAA→GAGTGGCA GAGTGGCTGA→GAGTGGCA GAGTGGCTTA→GAGTGGCA GAGTGGTCGA→GAGTGGTA GAGTGGTCTA→GAGTGGTA
GAGTGGTTCA→GAGTGGTA GAGTGGTTGA→GAGTGGTA GAGTGGTTTA→GAGTGGTA GCTAGGGAAG→GCTAGGAG GGACCCGTAT→GGACGTAT
GGTTTGGTTA→GGTTGGTA GTAAGTGGAA→GTAAGGAA GTAAGTGGTA→GTAAGGTA GTAGATGGTA→GTAGGGTA GTAGCGGTTA→GTAGGGTA

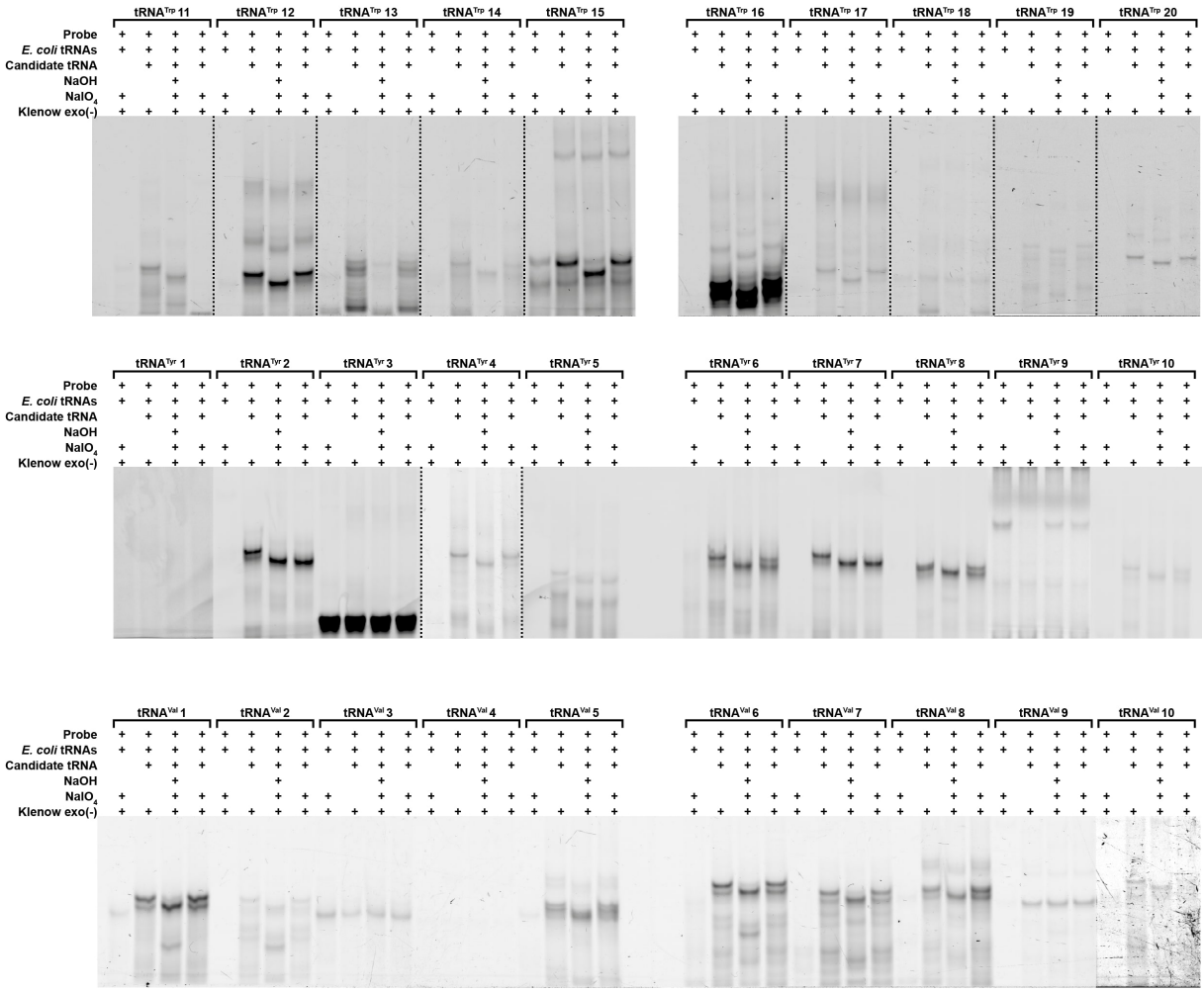
GTCAGGGGAG→GTCAGGGG
GTTAGGGGAG→GTTAGGGG
TAACTGGCTA→TAACGGCA
TAATTGGTTA→TAATGGTA
TAGCTGGTGC→TAGCGGTC
TAGTTGGTTA→TAGTGGTA
TATTGGTAAG→TATTGGTG
TCAGTTGGTA→TCAGGGTA
TGCCGGTAGG→TGCCGGTG
TGCTGGTTGC→TGCTGGTC
TGTTGGTCGC→TGTTGGTC
TTGCTGGTCG→TTGCGGTG
AACATGGCAGG→AACAGGCG
AACTGGAATTA→AACTGGAA
AACTTTGGTAG→AACTGGTG
AAGTGGTCATA→AAGTGGTA
AATATAGGTAT→AATAGGTT
AATCGGATTCA→AATCGGAA
AATGGTTTTAA→AAT-GGTA
AATTGGAATTA→AATTGGAA
AATTTCGGTAG→AATTGGTG
AATTTGGTTAG→AATTGGTG
ACTCGGTCTGT→ACTCGGTT
AGACTGGCCCA→AGACGGCA
AGATATGGTTA→AGATGGTA
AGCAAGGCTGA→AGCAGGCA
AGCAGGAAAGA→AGCAGGAA
AGCATGGTCAA→AGCAGGTA
AGCCAGGACTA→AGCCGGAA
AGCCAGGGCGA→AGCCGGGA
AGCCAGGTTAA→AGCCGGTA
AGCCCGGATCA→AGCCGGAA
AGCCCGGCTGA→AGCCGGCA
AGCCCGGTCTA→AGCCGGTA
AGCCTGGAATA→AGCCGGAA
AGCCTGGCAAA→AGCCGGCA
AGCCTGGCCTA→AGCCGGCA
AGCCTGGTCCA→AGCCGGTA
AGCCTGGTTGA→AGCCGGTA
AGCTAGGACAA→AGCTGGAA
AGCTAGGTCAA→AGCTGGTA
AGCTCGGCCAA→AGCTGGCA
AGCTCGGTCCA→AGCTGGTA
AGCTGCGGTTA→AGCTGGTA
AGCTGGCCTAG→AGCTGGCG
AGCTTGGACAA→AGCTGGAA
AGCTTGGCCCA→AGCTGGCA
AGCTTGGGTTA→AGCTGGGA
AGCTTGGTCTA→AGCTGGTA
AGGACGGTCGA→AGGAGGTA
AGGGGGGGATA→AGGGGGAA
AGTACGGCTTA→AGTAGGCA
AGTATGGCTAA→AGTAGGCA
AGTATGGTTGA→AGTAGGTA
AGTCAGGCCGT→AGTCGGCT
AGTCAGGCTTA→AGTCGGCA
AGTCAGGTCTA→AGTCGGTA
AGTCCGGACAA→AGTCGGAA
AGTCCGGCCGA→AGTCGGCA
AGTCCGGCTGA→AGTCGGCA
AGTCCGGTCTA→AGTCGGTA
AGTCGGGTCGA→AGTCGGTA
AGTCTGGATTA→AGTCGGAA
AGTCTGGCCTT→AGTCGGCT
AGTCTGGCTTA→AGTCGGCA
AGTCTGGTCTA→AGTCGGTA
AGTGGATTTCT→AGT-GGAT
AGTTAGGCTGA→AGTTGGCA
AGTTCGGCCTA→AGTTGGCA
AGTTCGGTCTA→AGTTGGTA
AGTTCTGGTTA→AGTTGGTA
AGTTGGGCTTA→AGTTGGGA
GTCCTGGTGG→GTCCGGTG
GTTCGGCTGA→GTTCGGCA
TAATGGTGTA→TAATGGTA
TAGAGGCCCA→TAGAGGCA
TAGCTGGTTA→TAGCGGTA
TATCCGGTGA→TATCGGTA
TATTGGTAGG→TATTGGTG
TCCCCGGTGG→TCCCGGTG
TGCTGGCACG→TGCTGGCG
TGTCGGCAAA→TGTCGGCA
TGTTGGTTGC→TGTTGGTC
TTGTTGGTCG→TTGTGGTG
AACCACGGTAG→AACCGGTG
AACTGGCCAGG→AACTGGCG
AAGCGGCCCAA→AAGCGGCA
AAGTGGTCTAA→AAGTGGTA
AATCAAGGTAG→AATCGGTG
AATCGGCTAGG→AATCGGCG
AATTCAGGTAG→AATTGGTG
AATTGGCAGAA→AATTGGCA
AATTTCGGTAT→AATTGGTT
AATTTTGGTTA→AATTGGTA
AGACAGGTCAA→AGACGGTA
AGACTGGCTGA→AGACGGCA
AGATGCGGTTA→AGATGGTA
AGCAAGGTCAA→AGCAGGTA
AGCAGGAAGAG→AGCAGGAG
AGCATGGTTGA→AGCAGGTA
AGCCAGGATTA→AGCCGGAA
AGCCAGGTCAA→AGCCGGTA
AGCCAGGTTGA→AGCCGGTA
AGCCCGGCCAA→AGCCGGCA
AGCCCGGCTTA→AGCCGGCA
AGCCCGGTTCA→AGCCGGTA
AGCCTGGACCA→AGCCGGAA
AGCCTGGCCAA→AGCCGGCA
AGCCTGGCTGA→AGCCGGCA
AGCCTGGTCGA→AGCCGGTA
AGCCTGGTTTA→AGCCGGTA
AGCTAGGATTA→AGCTGGAA
AGCTAGGTCCA→AGCTGGTA
AGCTCGGCCCA→AGCTGGCA
AGCTCGGTCGT→AGCTGGTT
AGCTGGCCAAG→AGCTGGCG
AGCTGGGCCGT→AGCTGGGT
AGCTTGGACTA→AGCTGGAA
AGCTTGGCCGA→AGCTGGCA
AGCTTGGTTAA→AGCTGGTA
AGGCCGGCCCA→AGGCGGCA
AGGTAGGTTTA→AGGTGGTA
AGTACGGTCTA→AGTAGGTA
AGTATGGCTGA→AGTAGGCA
AGTCAGGACGA→AGTCGGAA
AGTCAGGCCTA→AGTCGGCA
AGTCAGGTCAA→AGTCGGTA
AGTCAGGTTAA→AGTCGGTA
AGTCCGGATCA→AGTCGGAA
AGTCCGGCCGT→AGTCGGCT
AGTCCGGCTTA→AGTCGGCA
AGTCCGGTTCA→AGTCGGTA
AGTCNGGCTGA→AGTCGGCA
AGTCTGGCCAA→AGTCGGCA
AGTCTGGCTAA→AGTCGGCA
AGTCTGGTCAA→AGTCGGTA
AGTCTGGTTAA→AGTCGGTA
AGTGGGTAGTA→AGTGGGTA
AGTTAGGTTTA→AGTTGGTA
AGTTCGGCTGA→AGTTGGCA
AGTTCGGTTAA→AGTTGGTA
AGTTGCGGTTA→AGTTGCGA
AGTTGGTCTTA→AGTTGGTA
GTCTGGCTGA→GTCTGGCA
GTTCTGGTGT→GTTCGGTT
TAATGGTTCA→TAATGGTA
TAGAGGCCTA→TAGAGGCA
TAGTCTGGTA→TAGTGGTA
TATCGGCAGG→TATCGGCG
TCAGCTGGGA→TCAGGGGA
TCTCTGGTGA→TCTCGGTA
TGCTGGCAGG→TGCTGGCG
TGTCGGCAGG→TGTCGGCG
TTAACGGCAA→TTAAGGCA
TTGTTGGTTG→TTGTGGTG
AACCGGCTAGG→AACCGGCG
AACTTGGCTAA→AACTGGCA
AAGCGGCCGAA→AAGCGGCA
AAGTTGGCCAA→AAGTGGCA
AATCAGGCAGG→AATCGGCG
AATCGGTCTTA→AATCGGTA
AATTCGGCAGC→AATTGGCC
AATTCGGTCAT→AATTGGCT
AATTTGGACAA→AATTGGAA
ACAGCAGGCGA→ACAGGGCA
AGACCGGCCAA→AGACGGCA
AGACTGGCTTA→AGACGGCA
AGATGGATATA→AGATGGAA
AGCACGGCCAA→AGCAGGCA
AGCAGGATACA→AGCAGGAA
AGCATGGTTTA→AGCAGGTA
AGCCAGGCCAA→AGCCGGCA
AGCCAGGTCCA→AGCCGGTA
AGCCAGGTTTA→AGCCGGTA
AGCCCGGCCCA→AGCCGGCA
AGCCCGGTCAA→AGCCGGTA
AGCCCGGTTCT→AGCCGGTT
AGCCTGGACGA→AGCCGGAA
AGCCTGGCCCA→AGCCGGCA
AGCCTGGCTTA→AGCCGGCA
AGCCTGGTCGT→AGCCGGTT
AGCGGCCCTAA→AGC-GGCA
AGCTAGGCCAA→AGCTGGCA
AGCTAGGTCGA→AGCTGGTA
AGCTCGGCCTA→AGCTGGCA
AGCTCGGTTGA→AGCTGGTA
AGCTCGGTTCT→AGCTGGTT
AGCTGGCCGAA→AGCTGGCA
AGCTGGGCCGT→AGCTGGGT
AGCTGGTAGAG→AGCTGGTG
AGCTTGGATGA→AGCTGGAA
AGCTTGGCCTA→AGCTGGCA
AGCTTGGTCCA→AGCTGGTA
AGCTTGGTTGA→AGCTGGTA
AGGCCGGCTGA→AGGCGGCA
AGGTTGGTCTA→AGGTGGTA
AGTACGGTTGA→AGTAGGTA
AGTATGGTCCA→AGTAGGTA
AGTCAGGCCAA→AGTCGGCA
AGTCAGGCTAA→AGTCGGCA
AGTCAGGTCCA→AGTCGGTA
AGTCAGGTTGA→AGTCGGTA
AGTCCGGATTA→AGTCGGAA
AGTCCGGCCTA→AGTCGGCA
AGTCCGGTCAA→AGTCGGTA
AGTCCGGTTGA→AGTCGGTA
AGTCTCGGTTA→AGTCGGTA
AGTCTGGCCCA→AGTCGGCA
AGTCTGGCTAT→AGTCGGCT
AGTCTGGTCCA→AGTCGGTA
AGTCTGGTTCA→AGTCGGTA
AGTTAGGACAA→AGTTGGAA
AGTTCATGGTA→AGTTGGTA
AGTTCGGCTTA→AGTTGGCA
AGTTCGGTTGA→AGTTGGTA
AGTTGGCCAAG→AGTTGGCG
AGTTGGTTAGA→AGTTGGTA
GTGCTGGTAC→GTGCGGTC
TAAATGGCTA→TAAAGGCA
TAATTGGCTA→TAATGGCA
TAGCGGCCTA→TAGCGGCA
TAGTGGCCCA→TAGTGGCA
TATTGGCAGT→TATTGGCT
TCAGTAGGTA→TCAGGGTA
TGACGGTTTG→TGACGGTG
TGCTGGTAGG→TGCTGGTG
TGTCGGTAGG→TGTCGGTG
TTAACGGCTA→TTAAGGCA
TTTCGGTAGG→TTTCGGTG
AACCTTGGTAG→AACCGGTG
AACTTGGTCAG→AACTGGTG
AAGTAGGGTTA→AAGTGGTA
AAGTTTGGTAG→AAGTGGTG
AATCGGAATAA→AATCGGAA
AATCTGGCTAA→AATCGGCA
AATTCGGTAGC→AATTGGTC
AATTGGTTTAG→AATTGGTG
AATTTGGATGA→AATTGGAA
ACAGTTGGTTA→ACAGGGTA
AGACTGGACAA→AGACGGAA
AGATGGTTCTA→AGATGGTA
AGCACGGCTGA→AGCAGGCA
AGCATGGCCCA→AGCAGGCA
AGCCAGGACAA→AGCCGGAA
AGCCAGGCCCA→AGCCGGCA
AGCCAGGTCGA→AGCCGGTA
AGCCCGGACAA→AGCCGGAA
AGCCCGGCCGA→AGCCGGCA
AGCCCGGTCCA→AGCCGGTA
AGCCCGGTTGA→AGCCGGTA
AGCCTGGACTA→AGCCGGAA
AGCCTGGCTGA→AGCCGGCA
AGCCTGGTAGG→AGCCGGTG
AGCCTGGTCTA→AGCCGGTA
AGCGGCGTGGC→AGC-GGCC
AGCTAGGCTGA→AGCTGGCA
AGCTAGGTTAA→AGCTGGTA
AGCTCGGCTGA→AGCTGGCA
AGCTCGGTTTA→AGCTGGTA
AGCTGGCCGAG→AGCTGGCG
AGCTTAGGTTA→AGCTGGTA
AGCTTGGATTA→AGCTGGAA
AGCTTGGCTGA→AGCTGGCA
AGCTTGGTCGA→AGCTGGTA
AGCTTGGTTTA→AGCTGGTA
AGGCTGGATTA→AGGCGGAA
AGTAACGGATA→AGTAGGAA
AGTACGGTTTA→AGTAGGTA
AGTATGGTCGA→AGTAGGTA
AGTCAGGCCCA→AGTCGGCA
AGTCAGGCTCA→AGTCGGCA
AGTCAGGTCGA→AGTCGGTA
AGTCAGGTTTA→AGTCGGTA
AGTCCGGCCAA→AGTCGGCA
AGTCCGGCTAA→AGTCGGCA
AGTCCGGTCCA→AGTCGGTA
AGTCCGGTTGT→AGTCGGTT
AGTCTGGACAA→AGTCGGAA
AGTCTGGCCGA→AGTCGGCA
AGTCTGGCTCA→AGTCGGCA
AGTCTGGTCGA→AGTCGGTA
AGTCTGGTTGA→AGTCGGTA
AGTTAGGCCCA→AGTTGGCA
AGTTCCGGTCA→AGTTGGTA
AGTTCGGTCGA→AGTTGGTA
AGTTCGGTCCA→AGTTGGTA
AGTTCGGTTTA→AGTTGGTA
AGTTGGCCAAG→AGTTGGCG
AGTTGGTTGAA→AGTTGGTA
GTTAGGGAAG→GTTAGGAG
TAACTGGATA→TAACGGAA
TAATTGGTAA→TAATGGTA
TAGCGGTCCA→TAGCGGTA
TAGTTGGCTA→TAGTGGCA
TATTGGCGAA→TATTGGCA
TCAGTCGGTA→TCAGGGTA
TGCCGGCAGG→TGCCGGCG
TGCTGGTATG→TGCTGGTG
TGTTGGCAGG→TGTTGGCG
TTCCGGCAGG→TTCCGGCG
AACACTGGTAG→AACAGGTG
AACTCAGGTAG→AACTGGTG
AACTTGGTTAA→AACTGGTA
AAGTGGCCTAC→AAGTGGCC
AATAGGTTTAG→AATAGGTG
AATCGGAATTA→AATCGGAA
AATGCAGGTAG→AATGGGTG
AATTCTGGTAG→AATTGGTG
AATTTCGGCAG→AATTGGCG
AATTTGGCAGC→AATTGGCC
ACGCGTGGTTA→ACGCGGTA
AGACTGGCCAA→AGACGGCA
AGACTGGTCCA→AGACGGTA
AGATTGGCCCA→AGATGGCA
AGCACGGTCAA→AGCAGGTA
AGCATGGCTGA→AGCAGGCA
AGCCAGGACCA→AGCCGGAA
AGCCAGGCCTA→AGCCGGCA
AGCCAGGTCTA→AGCCGGTA
AGCCCGGACTA→AGCCGGAA
AGCCCGGCCTA→AGCCGGCA
AGCCCGGTCGA→AGCCGGTA
AGCCCGGTTTA→AGCCGGTA
AGCCTGGATTA→AGCCGGAA
AGCCTGGCCGT→AGCCGGCT
AGCCTGGTCAA→AGCCGGTA
AGCCTGGTTAA→AGCCGGTA
AGCGGTAGAGC→AGC-GGTC
AGCTAGGGCAA→AGCTGGGA
AGCTCAGGTTA→AGCTGGTA
AGCTCGGTCAA→AGCTGGTA
AGCTCTGGTTA→AGCTGGTA
AGCTTCGGTTA→AGCTGGTA
AGCTTGGCCAA→AGCTGGCA
AGCTTGGCTTA→AGCTGGCA
AGCTTGGTCGT→AGCTGGTT
AGCTTTGGTTA→AGCTGGTA
AGGCTGGTCTA→AGGCGGTA
AGTACGGCTGA→AGTAGGCA
AGTATGGCCAA→AGTAGGCA
AGTATGGTTAA→AGTAGGTA
AGTCAGGCCGA→AGTCGGCA
AGTCAGGCTGA→AGTCGGCA
AGTCAGGTCGT→AGTCGGTT
AGTCCGCCTGA→AGTCGCCA
AGTCCGGCCCA→AGTCGGCA
AGTCCGGCTCA→AGTCGGCA
AGTCCGGTCGA→AGTCGGTA
AGTCCGGTTTA→AGTCGGTA
AGTCTGGACTA→AGTCGGAA
AGTCTGGCCTA→AGTCGGCA
AGTCTGGCTGA→AGTCGGCA
AGTCTGGTCGT→AGTCGGTT
AGTCTGGTTTA→AGTCGGTA
AGTTAGGCCTA→AGTTGGCA
AGTTCGGCCGT→AGTTGGCT
AGTTCGGTCGT→AGTTGGTT
AGTTCTGGTCA→AGTTGGTA
AGTTGGCCGTA→AGTTGGCA
AGTTGGTTTAG→AGTTGGTG

Chapter V – Appendix

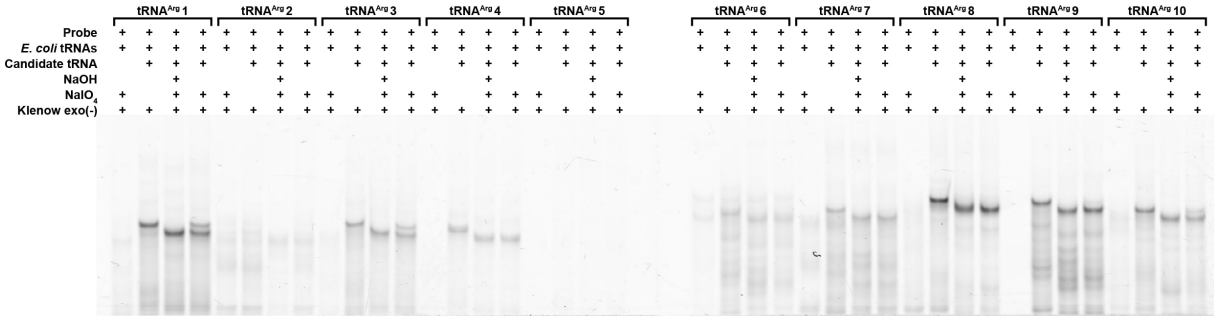
AGTTTAGGTCA→AGTTGGTA	AGTTTCGGTTA→AGTTGGTA	AGTTTGGCCAA→AGTTGGCA	AGTTTGGCCGA→AGTTGGCA	AGTTTGGCCGT→AGTTGGCT
AGTTTGGCTGA→AGTTGGCA	AGTTTGGCTTA→AGTTGGCA	AGTTTGGTACA→AGTTGGTA	AGTTTGGTCAA→AGTTGGTA	AGTTTGGTCCA→AGTTGGTA
AGTTTGGTCGA→AGTTGGTA	AGTTTGGTCTA→AGTTGGTA	AGTTTGGTTAA→AGTTGGTA	AGTTTGGTTGA→AGTTGGTA	AGTTTGGTTTA→AGTTGGTA
AGTTTTGGTAA→AGTTGGTA	AGTTTTGGTTA→AGTTGGTA	ATACGGCTTGG→ATACGGCG	ATACTGGTTGG→ATACGGTG	ATACTGGTTTG→ATACGGTG
ATATTGGTTTG→ATATGGTG	ATCTTGGTAGA→ATCTGGTA	ATGCATGGTTA→ATGCGGTA	ATGCCTGGCAA→ATGCGGCA	ATGCCTGGTCA→ATGCGGTA
ATGCCTGGTTA→ATGCGGTA	ATGCTTGGTTA→ATGCGGTA	ATGTTGGTTTG→ATGTGGTG	CAAGTGGCTGA→CAAGGGCA	CAATCTGGCAG→CAATGGCG
CAGCCGGTTAG→CAGCGGTG	CAGGTGGTTAG→CAGGGGTG	CAGTTGGTTAG→CAGTGGTG	CATGGCGGTCT→CATGGGTT	CCGAGTGGCTG→CCGAGGCG
CGAGCGGCCAA→CGAGGGCA	CGAGTGGCCAA→CGAGGGCA	CGAGTGGCTGA→CGAGGGCA	CGCAGATTGAC→CGCAGATC	CGCTCGGAGAG→CGCTGGAG
CTATTGGTATG→CTATGGTG	CTCAGCTGGGA→CTCAGGGA	CTCAGTTGGTA→CTCAGGTA	GAAATTGGTAG→GAAAGGTG	GAAGAGGCTAA→GAAGGGCA
GAAGTGGTTTA→GAAGGGTA	GAGCGGCCAAA→GAGCGGCA	GAGCGGTTTAT→GAGCGGTT	GAGCTGGTTTA→GAGCGGTA	GAGTGGCTGAA→GAGTGGCA
GAGTGGCTTAT→GAGTGGCT	GAGTGGTCGAA→GAGTGGTA	GAGTGGTTTAT→GAGTGGTT	GAGTTGGTTTA→GAGTGGTA	GCAATTGGATA→GCAAGGAA
GCAGAGGCCGC→GCAGGGCC	GCAGCCTGGTA→GCAGGGTA	GCAGCTTGGTA→GCAGGGTA	GCAGTTCGGTA→GCAGGGTA	GGAATTGGCAG→GGAAGGCG
GGAATTGGTAG→GGAAGGTG	GGCTTGGTAGC→GGCTGGTC	GTAGGGGTAGC→GTAGGGTC	GTGCTGGTTTC→GTGCGGTC	GTGTTGGTCAA→GTGTGGTA
GTGTTGGTTTC→GTGTGGTC	TAGCGGCCCAA→TAGCGGCA	TAGGAGGCCCA→TAGGGGCA	TAGTCCTGGTA→TAGTGGTA	TAGTGGCCTAG→TAGTGGCG
TATATAGGTTA→TATAGGTA	TATCTGGTGAT→TATCGGTT	TCAGCTGGATA→TCAGGGAA	TCAGCTGGTTA→TCAGGGTA	TCAGGAATAGC→TCA-GGAC
TCAGGTGGTTA→TCAGGGTA	TCAGTTGGTAG→TCAGGGTG	TCAGTTGGTTA→TCAGGGTA	TGACGGTGGTA→TGACGGTA	TGAGTGGTTGA→TGAGGGTA
TGTAGCGGTTA→TGTAGGTA	TGTTGGGCAAG→TGTTGGCG	TTAACGGTTTA→TTAAGGTA	TTGTTGGTATG→TTGTGGTG	TTGTTGGTCAA→TTGTGGTA
ACTCAGTTGGGA→ACTCGTTA	AGTCCGGCCGTA→AGTCGGCA	AGTCCGGCTGTA→AGTCGGCA	AGTCCGGTCGTA→AGTCGGTA	TAGCACTGTGGGA→TAGCGTGA

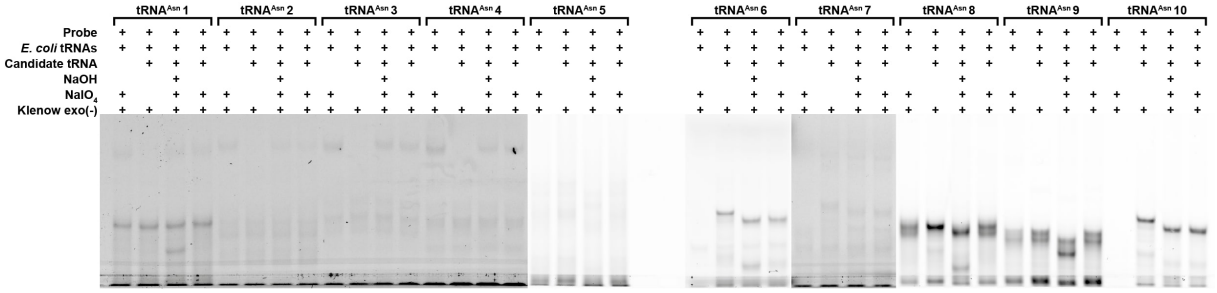
tREX Screening Gels

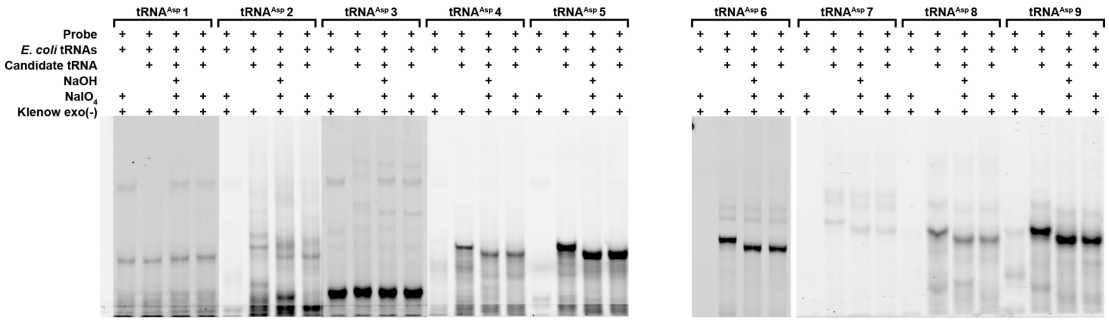
Complete set of tREX screening gels.

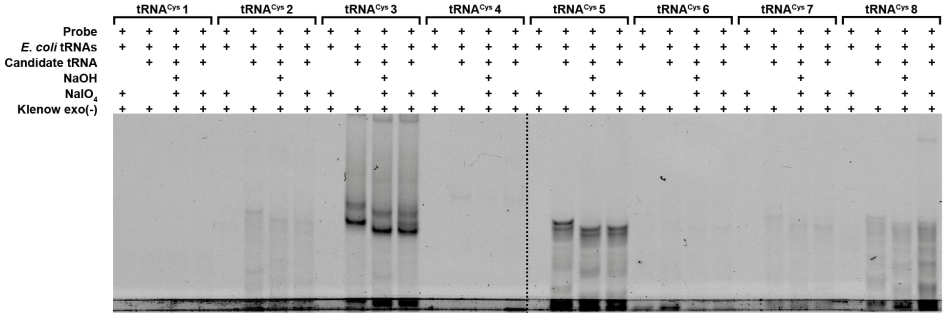


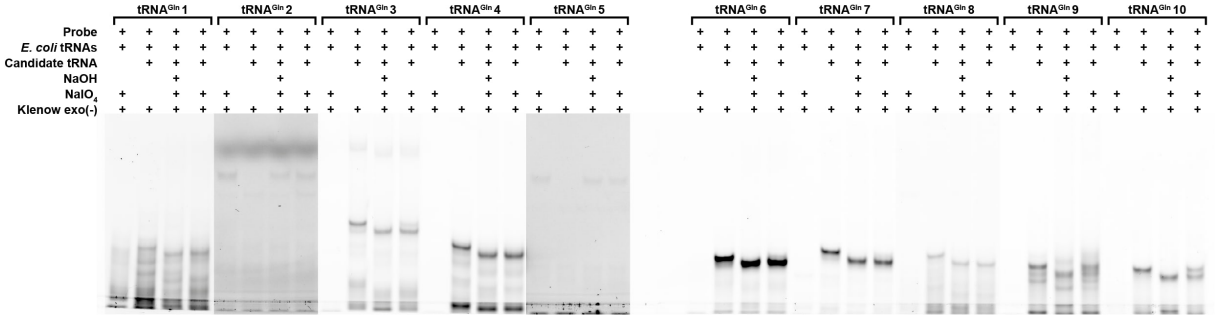
Chapter V – Appendix

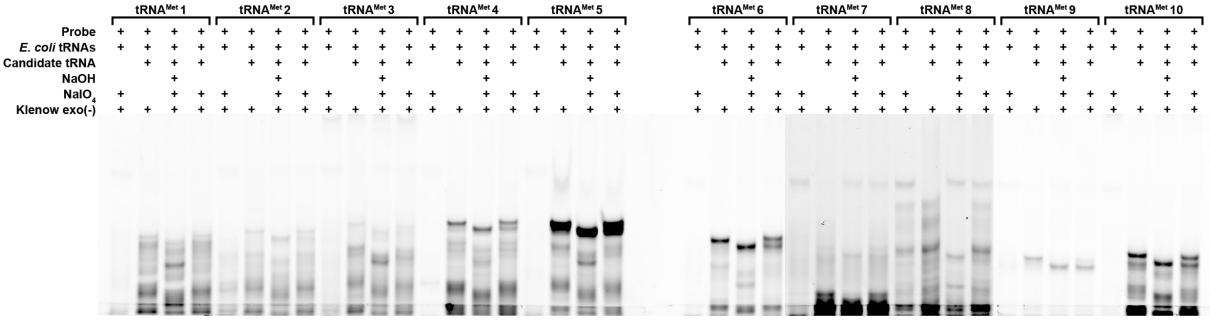


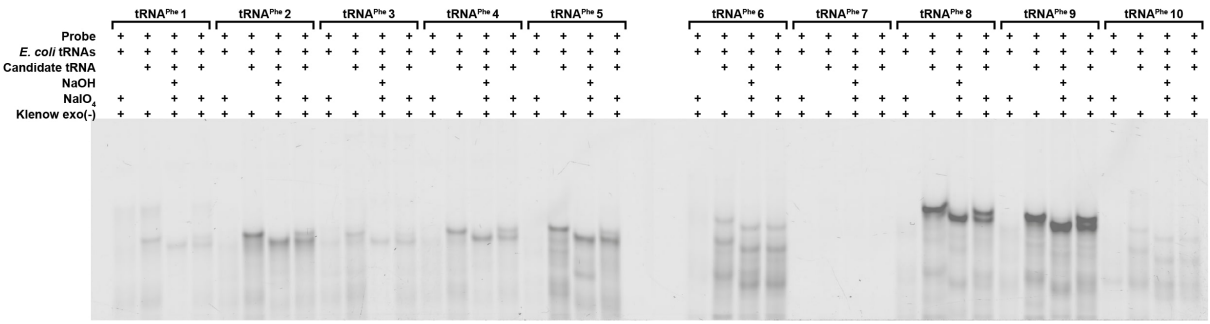


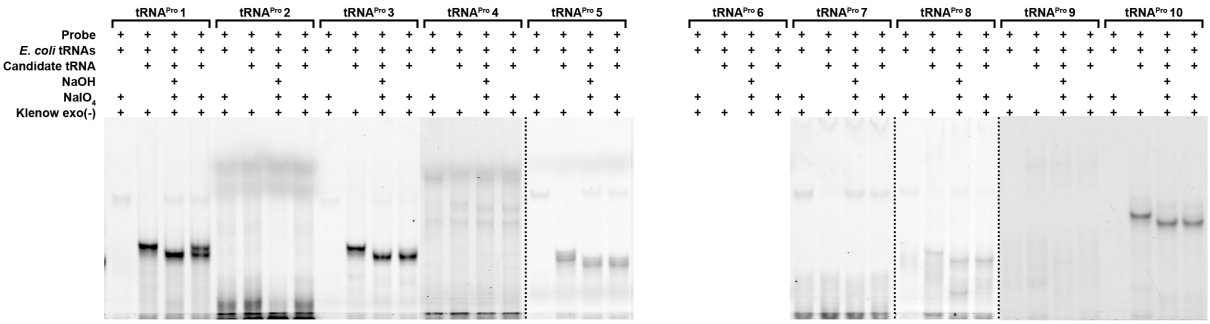


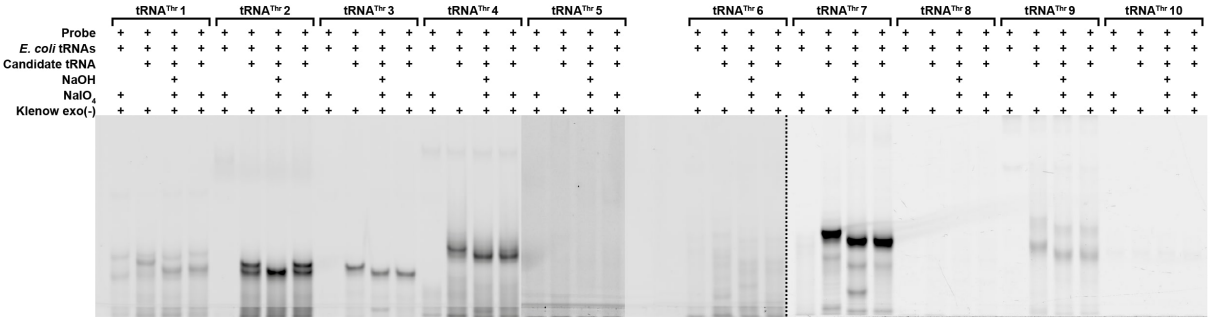


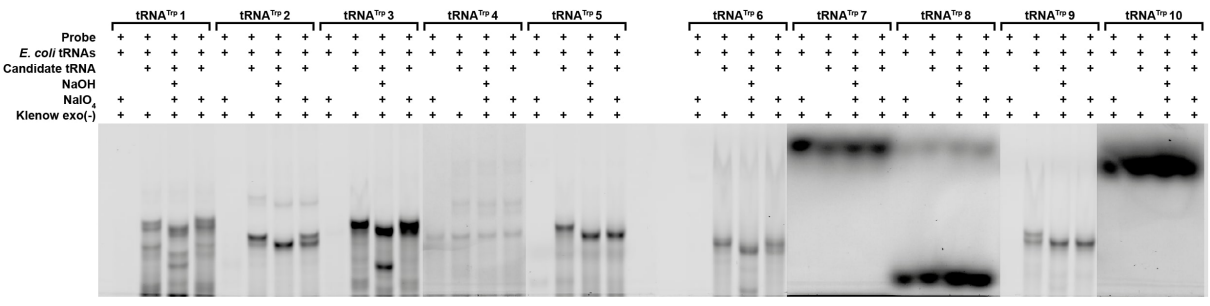



tRNA^Met 1
tRNA^Met 2
tRNA^Met 3
tRNA^Met 4
tRNA^Met 5
tRNA^Met 6
tRNA^Met 7
tRNA^Met 8
tRNA^Met 9
tRNA^Met 10
Probe
E. coli tRNAs
Candidate tRNA
NaOH
NaIO4
Klenow exo(-)

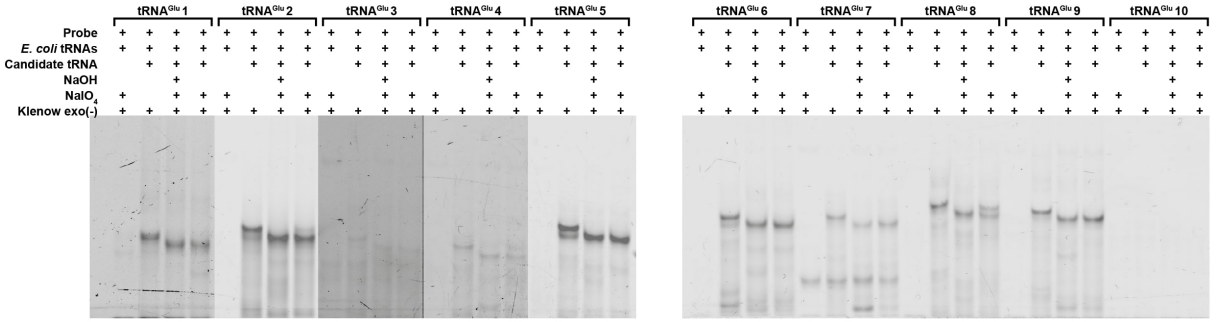

tRNA^Phe 1
tRNA^Phe 2
tRNA^Phe 3
tRNA^Phe 4
tRNA^Phe 5
tRNA^Phe 6
tRNA^Phe 7
tRNA^Phe 8
tRNA^Phe 9
tRNA^Phe 10
Probe
E. coli tRNAs
Candidate tRNA
NaOH
NaIO4
Klenow exo(-)

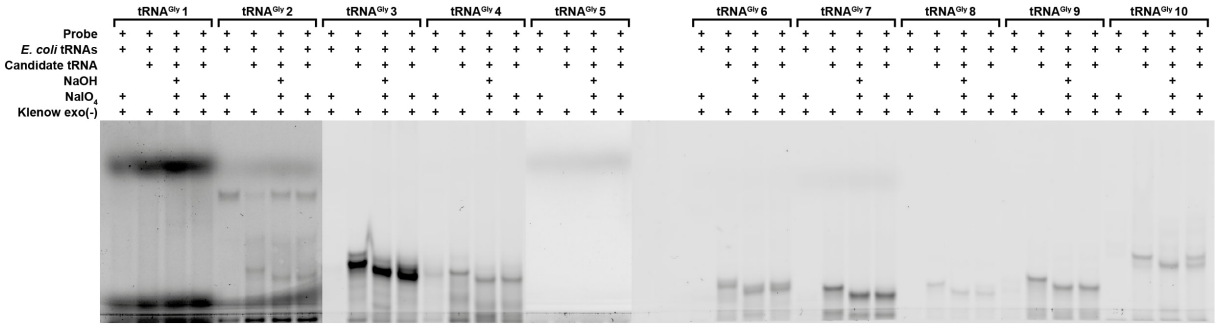

tRNA^Pro 1
tRNA^Pro 2
tRNA^Pro 3
tRNA^Pro 4
tRNA^Pro 5
tRNA^Pro 6
tRNA^Pro 7
tRNA^Pro 8
tRNA^Pro 9
tRNA^Pro 10
Probe
E. coli tRNAs
Candidate tRNA
NaOH
NaIO4
Klenow exo(-)

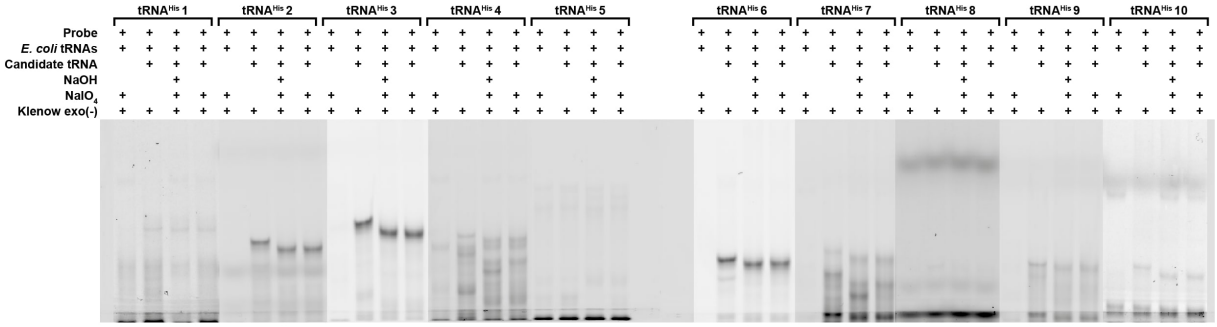

tRNA^Thr 1
tRNA^Thr 2
tRNA^Thr 3
tRNA^Thr 4
tRNA^Thr 5
tRNA^Thr 6
tRNA^Thr 7
tRNA^Thr 8
tRNA^Thr 9
tRNA^Thr 10
Probe
E. coli tRNAs
Candidate tRNA
NaOH
NaIO4
Klenow exo(-)

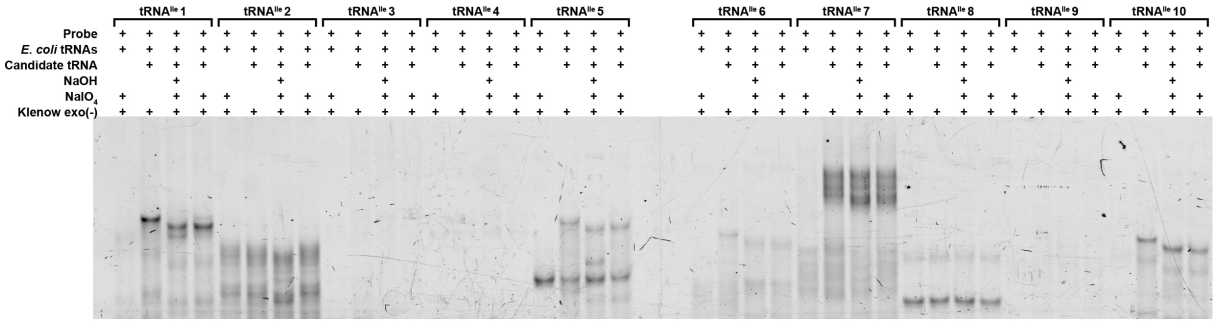

tRNA^Trp 1
tRNA^Trp 2
tRNA^Trp 3
tRNA^Trp 4
tRNA^Trp 5
tRNA^Trp 6
tRNA^Trp 7
tRNA^Trp 8
tRNA^Trp 9
tRNA^Trp 10
Probe
E. coli tRNAs
Candidate tRNA
NaOH
NaIO4
Klenow exo(-)

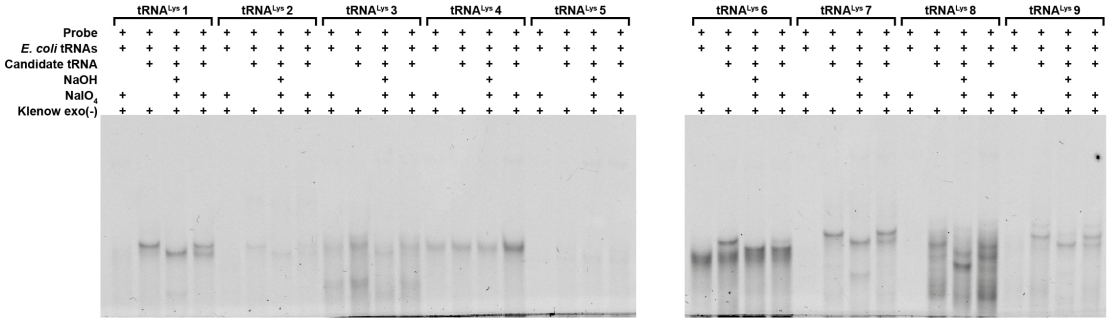
Chapter V – Appendix

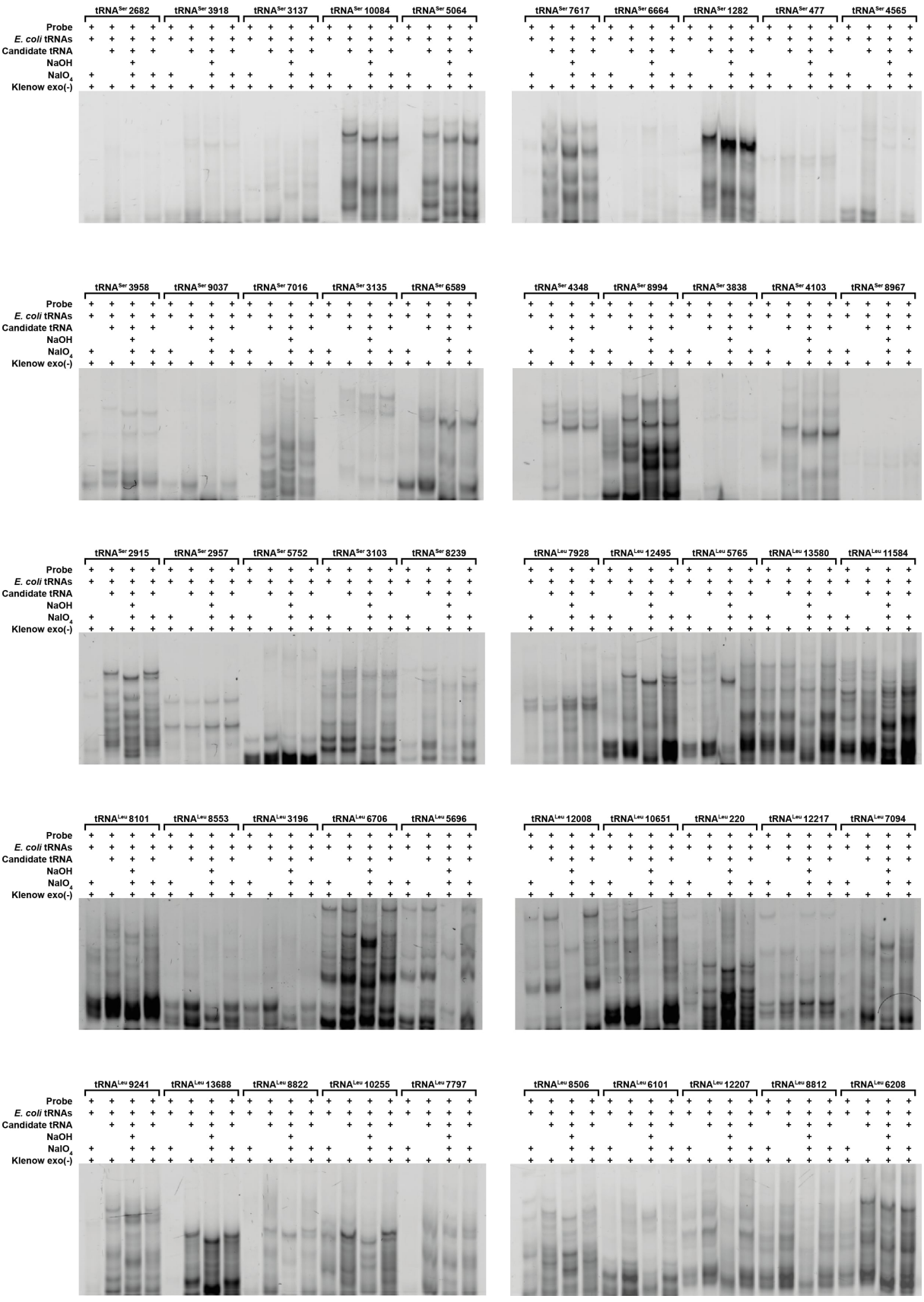











tRNA^{Ser} 2682
tRNA^{Ser} 3918
tRNA^{Ser} 3137
tRNA^{Ser} 10084
tRNA^{Ser} 5064
tRNA^{Ser} 7617
tRNA^{Ser} 6664
tRNA^{Ser} 1282
tRNA^{Ser} 477
tRNA^{Ser} 4565
tRNA^{Ser} 3958
tRNA^{Ser} 9037
tRNA^{Ser} 7016
tRNA^{Ser} 3135
tRNA^{Ser} 6589
tRNA^{Ser} 4348
tRNA^{Ser} 8994
tRNA^{Ser} 3838
tRNA^{Ser} 4103
tRNA^{Ser} 8967
tRNA^{Ser} 2915
tRNA^{Ser} 2957
tRNA^{Ser} 5752
tRNA^{Ser} 3103
tRNA^{Ser} 8239
tRNA^{Leu} 7928
tRNA^{Leu} 12495
tRNA^{Leu} 5765
tRNA^{Leu} 13580
tRNA^{Leu} 11584
tRNA^{Leu} 8101
tRNA^{Leu} 8553
tRNA^{Leu} 3196
tRNA^{Leu} 6706
tRNA^{Leu} 5696
tRNA^{Leu} 12008
tRNA^{Leu} 10651
tRNA^{Leu} 220
tRNA^{Leu} 12217
tRNA^{Leu} 7094
tRNA^{Leu} 9241
tRNA^{Leu} 13688
tRNA^{Leu} 8822
tRNA^{Leu} 10255
tRNA^{Leu} 7797
tRNA^{Leu} 8506
tRNA^{Leu} 6101
tRNA^{Leu} 12207
tRNA^{Leu} 8812
tRNA^{Leu} 6208
Probe
E. coli tRNAs
Candidate tRNA
NaOH
$NaIO_4$
Klenow exo(-)

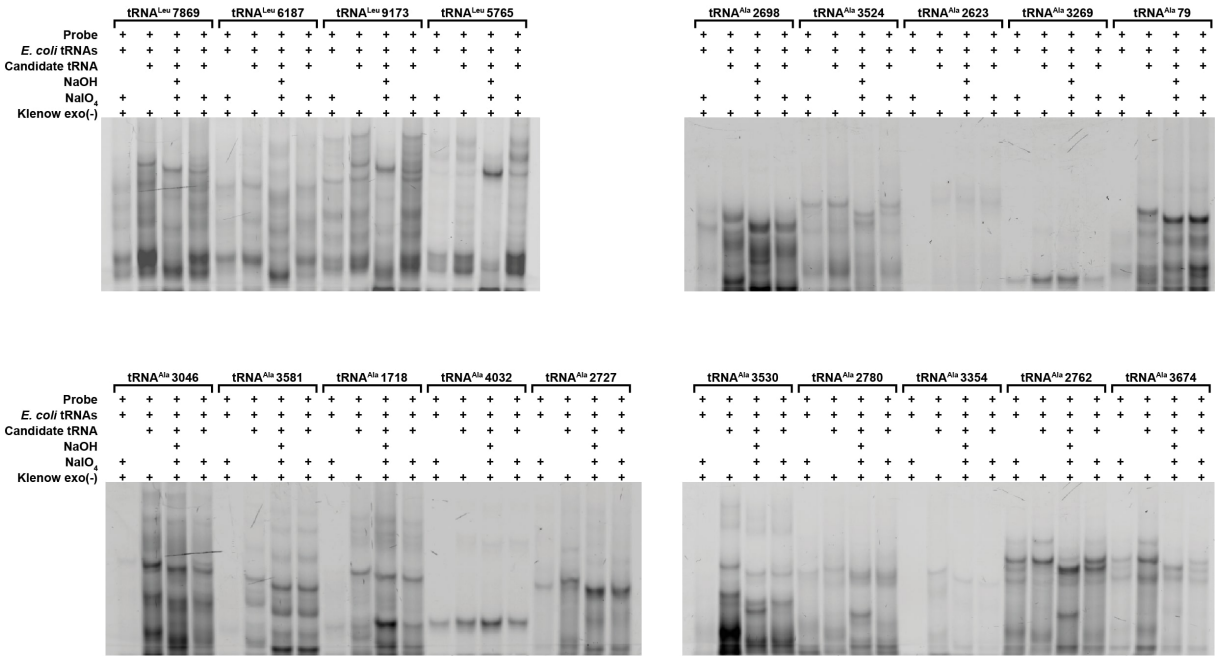
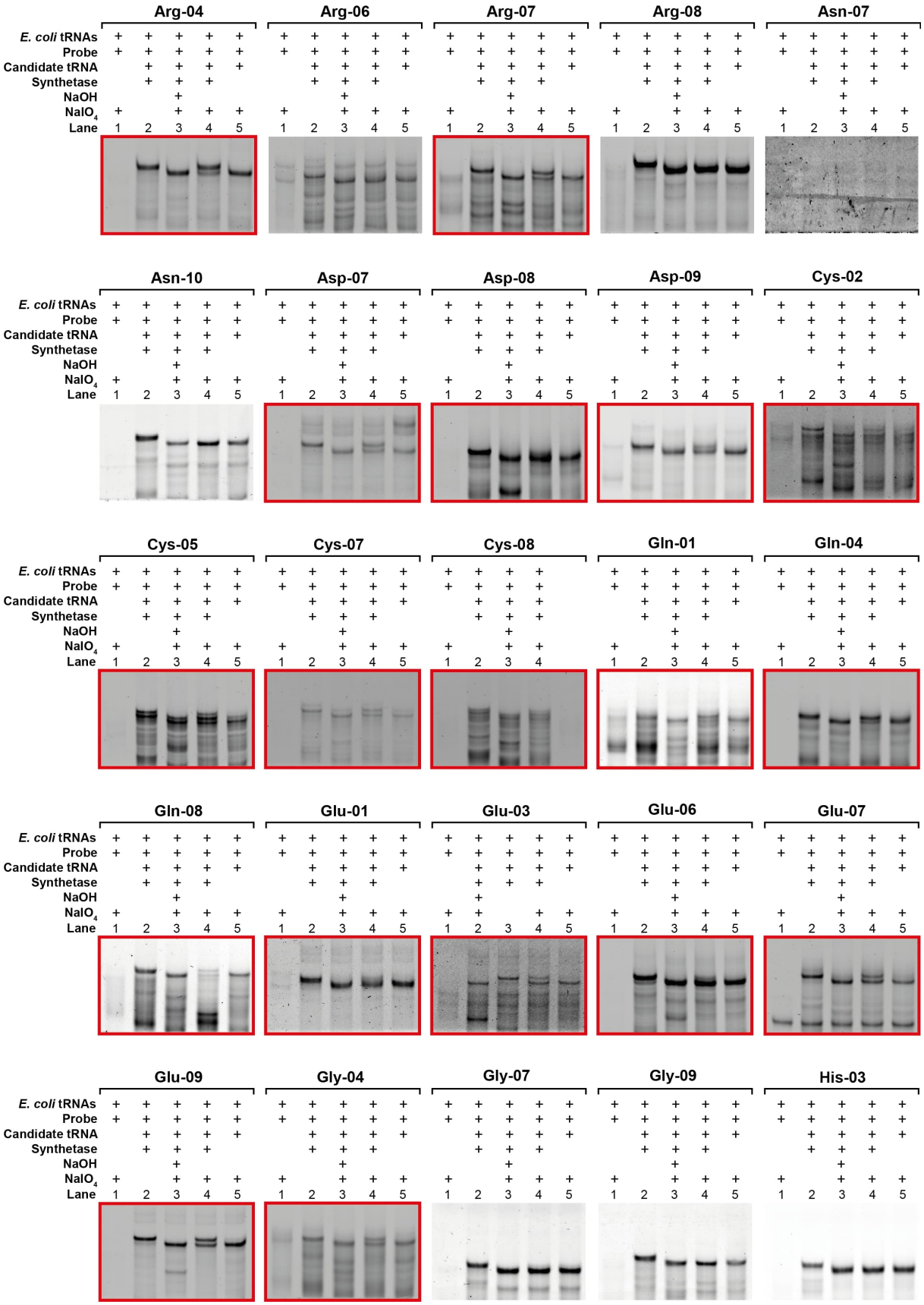
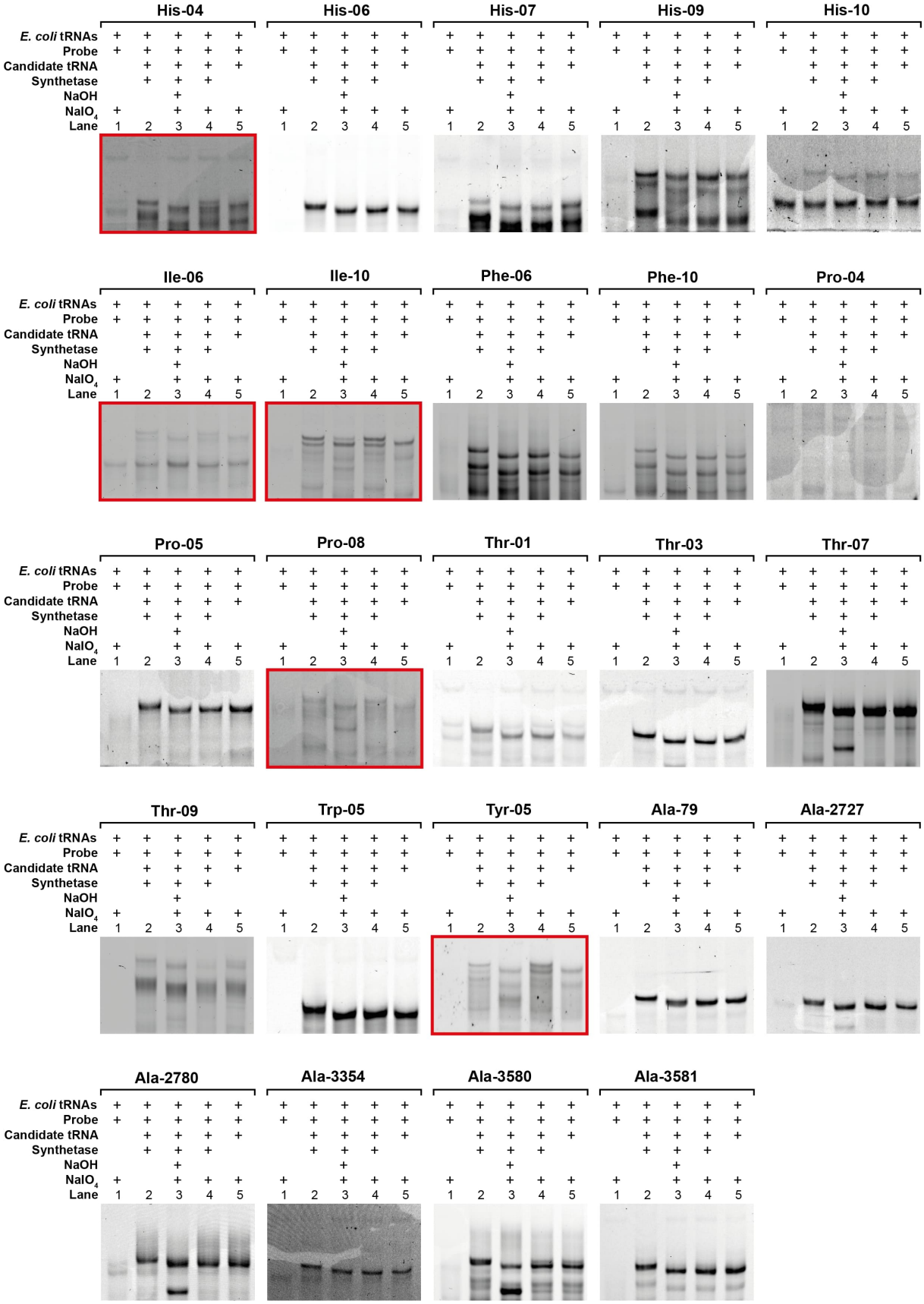


Figure 5.1*: Complete set of gels for tREX screening of tRNA orthogonality in E. coli. For each tRNA, a specific DNA probe was designed to be complementary to it. Furthermore, the specificity of each probe was tested on a tRNA extract from wild type E. coli DH10b. For each tRNA under investigation, a control for the electrophoretic mobility of the extended species was generated by omitting the $NaIO_4$ oxidation, while a control for the electrophoretic mobility of the unextended species was generated by performing $NaIO_4$ oxidation following chemical deacylation by NaOH. Orthogonality was assessed by verifying if the oxidised tRNA sample displayed the a band with the same electrophoretic mobility as the control for the unextended species and no bands with the same electrophoretic mobility as the control for the extended species.*


Arg-04
Arg-06
Arg-07
Arg-08
Asn-07
Asn-10
Asp-07
Asp-08
Asp-09
Cys-02
Cys-05
Cys-07
Cys-08
Gln-01
Gln-04
Gln-08
Glu-01
Glu-03
Glu-06
Glu-07
Glu-09
Gly-04
Gly-07
Gly-09
His-03
E. coli tRNAs
Probe
Candidate tRNA
Synthetase
NaOH
NaIO4
Lane

Chapter V – Appendix



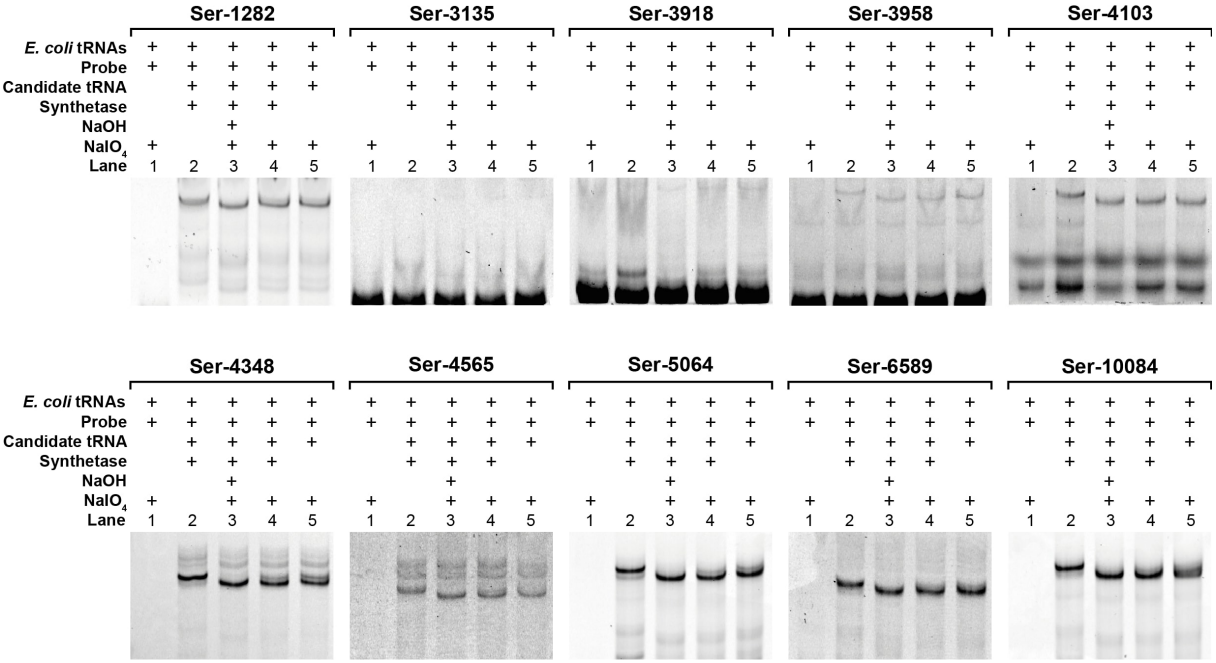


Figure 5.2*: Complete set of gels containing the results of the tREX screening for the aminoacyltion of the orthogonal tRNA by their cognate synthetase. For each tRNA under investigation, a control for the electrophoretic mobility of the extended and unextended species was generated as described before. The aminoacylation status of the tRNA was verified when it was co-expressed in E. coli cells together with its cognate synthetase from the same organism. Samples for which a synthetase-dependent aminoacylation was observed are highlighted by a red box.*

PrimDesign

Below the code, written in Mathematica, of the computational tool used to design primer for site-saturation libraries which include only the amino acids of choice.

Interface for Translational Table Editing

```
Interpretation[
{ttt1="Phe",ttc1="Phe",tta1="Leu",ttg1="Leu",tct1="Ser",tcc1="Ser",tca1="Ser",tcg1="Ser",taa1="***",tat1="T
yr",tac1="Tyr",tag1="***",tga1="***",tgt1="Cys",tgc1="Cys",tgg1="Trp",ctt1="Leu",ctc1="Leu",cta1="Leu",ctg1
="Leu",cct1="Pro",ccc1="Pro",cca1="Pro",ccg1="Pro",cat1="His",cac1="His",caa1="Gln",cag1="Gln",cgt1="Arg",c
gc1="Arg",cga1="Arg",cgg1="Arg",att1="Ile",atc1="Ile",ata1="Ile",atg1="Met",act1="Thr",acc1="Thr",aca1="Thr
",acg1="Thr",aat1="Asn",aac1="Asn",aaa1="Lys",aag1="Lys",agt1="Ser",agc1="Ser",aga1="Arg",agg1="Arg",gtt1="
Val",gtc1="Val",gta1="Val",gtg1="Val",gct1="Ala",gcc1="Ala",gca1="Ala",gcg1="Ala",gat1="Asp",gac1="Asp",gaa
1="Glu",gag1="Glu",ggt1="Gly",ggc1="Gly",gga1="Gly",ggg1="Gly",ttt="***",ttc="***",tta="***",ttg="***",tct=
"***",tcc="***",tca="***",tcg="***",taa="taa",tat="***",tac="***",tag="tag",tga="tga",tgt="***",tgc="***",t
gg="***",ctt="***",ctc="***",cta="cta",ctg="***",cct="***",ccc="ccc",cca="***",ccg="***",cat="***",cac="***
",caa="***",cag="***",cgt="***",cgc="***",cga="cga",cgg="cgg",att="***",atc="***",ata="ata",atg="***",act="
***",acc="***",aca="***",acg="***",aat="***",aac="***",aaa="***",aag="***",agt="***",agc="***",aga="aga",ag
g="agg",gtt="***",gtc="***",gta="***",gtg="***",gct="***",gcc="***",gca="***",gcg="***",gat="***",gac="***"
,gaa="***",gag="***",ggt="***",ggc="***",gga="gga",ggg="***"},
 Button["Edit Translational Table",
  CreateDialog[{
    Panel@Grid[{
       {Style["Tick the codons you want to exlude from the
calculation",Bold,Medium],SpanFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft,Spa
nFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft},{Checkbox[Dynamic@ttt,
{"***","ttt"}],Style["ttt",Bold],InputField[Dynamic@ttt1,String,FieldSize->3],Checkbox[Dynamic@tct,
{"***","tct"}],Style["tct",Bold],InputField[Dynamic@tct1,String,FieldSize->3],Checkbox[Dynamic@tat,
{"***","tat"}],Style["tat",Bold],InputField[Dynamic@tat1,String,FieldSize->3],Checkbox[Dynamic@tgt,
{"***","tgt"}],Style["tgt",Bold],InputField[Dynamic@tgt1,String,FieldSize->3]},
       {Checkbox[Dynamic@ttc,{"***","ttc"}],Style["ttc",Bold],InputField[Dynamic@ttc1,String,FieldSize-
>3],Checkbox[Dynamic@tcc,{"***","tcc"}],Style["tcc",Bold],InputField[Dynamic@tcc1,String,FieldSize-
>3],Checkbox[Dynamic@tac,{"***","tac"}],Style["tac",Bold],InputField[Dynamic@tac1,String,FieldSize-
>3],Checkbox[Dynamic@tgc,{"***","tgc"}],Style["tgc",Bold],InputField[Dynamic@tgc1,String,FieldSize->3]},
       {Checkbox[Dynamic@tta,{"***","tta"}],Style["tta",Bold],InputField[Dynamic@tta1,String,FieldSize-
>3],Checkbox[Dynamic@tca,{"***","tca"}],Style["tca",Bold],InputField[Dynamic@tca1,String,FieldSize-
>3],Checkbox[Dynamic@taa,{"***","taa"}],Style["taa",Bold],InputField[Dynamic@taa1,String,FieldSize-
>3],Checkbox[Dynamic@tga,{"***","tga"}],Style["tga",Bold],InputField[Dynamic@tga1,String,FieldSize->3]},
       {Checkbox[Dynamic@ttg,{"***","ttg"}],Style["ttg",Bold],InputField[Dynamic@ttg1,String,FieldSize-
>3],Checkbox[Dynamic@tcg,{"***","tcg"}],Style["tcg",Bold],InputField[Dynamic@tcg1,String,FieldSize-
>3],Checkbox[Dynamic@tag,{"***","tag"}],Style["tag",Bold],InputField[Dynamic@tag1,String,FieldSize-
>3],Checkbox[Dynamic@tgg,{"***","tgg"}],Style["tgg",Bold],InputField[Dynamic@tgg1,String,FieldSize->3]},
       {Checkbox[Dynamic@ctt,{"***","ctt"}],Style["ctt",Bold],InputField[Dynamic@ctt1,String,FieldSize-
>3],Checkbox[Dynamic@cct,{"***","cct"}],Style["cct",Bold],InputField[Dynamic@cct1,String,FieldSize-
>3],Checkbox[Dynamic@cat,{"***","cat"}],Style["cat",Bold],InputField[Dynamic@cat1,String,FieldSize-
>3],Checkbox[Dynamic@cgt,{"***","cgt"}],Style["cgt",Bold],InputField[Dynamic@cgt1,String,FieldSize->3]},
       {Checkbox[Dynamic@ctc,{"***","ctc"}],Style["ctc",Bold],InputField[Dynamic@ctc1,String,FieldSize-
>3],Checkbox[Dynamic@ccc,{"***","ccc"}],Style["ccc",Bold],InputField[Dynamic@ccc1,String,FieldSize-
>3],Checkbox[Dynamic@cac,{"***","cac"}],Style["cac",Bold],InputField[Dynamic@cac1,String,FieldSize-
>3],Checkbox[Dynamic@cgc,{"***","cgc"}],Style["cgc",Bold],InputField[Dynamic@cgc1,String,FieldSize->3]},
       {Checkbox[Dynamic@cta,{"***","cta"}],Style["cta",Bold],InputField[Dynamic@cta1,String,FieldSize-
>3],Checkbox[Dynamic@cca,{"***","cca"}],Style["cca",Bold],InputField[Dynamic@cca1,String,FieldSize-
>3],Checkbox[Dynamic@caa,{"***","caa"}],Style["caa",Bold],InputField[Dynamic@caa1,String,FieldSize-
>3],Checkbox[Dynamic@cga,{"***","cga"}],Style["cga",Bold],InputField[Dynamic@cga1,String,FieldSize->3]},
       {Checkbox[Dynamic@ctg,{"***","ctg"}],Style["ctg",Bold],InputField[Dynamic@ctg1,String,FieldSize-
>3],Checkbox[Dynamic@ccg,{"***","ccg"}],Style["ccg",Bold],InputField[Dynamic@ccg1,String,FieldSize-
>3],Checkbox[Dynamic@cag,{"***","cag"}],Style["cag",Bold],InputField[Dynamic@cag1,String,FieldSize-
>3],Checkbox[Dynamic@cgg,{"***","cgg"}],Style["cgg",Bold],InputField[Dynamic@cgg1,String,FieldSize->3]},
       {Checkbox[Dynamic@att,{"***","att"}],Style["att",Bold],InputField[Dynamic@att1,String,FieldSize-
>3],Checkbox[Dynamic@act,{"***","act"}],Style["act",Bold],InputField[Dynamic@act1,String,FieldSize-
```

```
>3],Checkbox[Dynamic@aat,{"***","aat"}],Style["aat",Bold],InputField[Dynamic@aat1,String,FieldSize-
>3],Checkbox[Dynamic@agt,{"***","agt"}],Style["agt",Bold],InputField[Dynamic@agt1,String,FieldSize->3]},
       {Checkbox[Dynamic@atc,{"***","atc"}],Style["atc",Bold],InputField[Dynamic@atc1,String,FieldSize-
>3],Checkbox[Dynamic@acc,{"***","acc"}],Style["acc",Bold],InputField[Dynamic@acc1,String,FieldSize-
>3],Checkbox[Dynamic@aac,{"***","aac"}],Style["aac",Bold],InputField[Dynamic@aac1,String,FieldSize-
>3],Checkbox[Dynamic@agc,{"***","agc"}],Style["agc",Bold],InputField[Dynamic@agc1,String,FieldSize->3]},
       {Checkbox[Dynamic@ata,{"***","ata"}],Style["ata",Bold],InputField[Dynamic@ata1,String,FieldSize-
>3],Checkbox[Dynamic@aca,{"***","aca"}],Style["aca",Bold],InputField[Dynamic@aca1,String,FieldSize-
>3],Checkbox[Dynamic@aaa,{"***","aaa"}],Style["aaa",Bold],InputField[Dynamic@aaa1,String,FieldSize-
>3],Checkbox[Dynamic@aga,{"***","aga"}],Style["aga",Bold],InputField[Dynamic@aga1,String,FieldSize->3]},
       {Checkbox[Dynamic@atg,{"***","atg"}],Style["atg",Bold],InputField[Dynamic@atg1,String,FieldSize-
>3],Checkbox[Dynamic@acg,{"***","acg"}],Style["acg",Bold],InputField[Dynamic@acg1,String,FieldSize-
>3],Checkbox[Dynamic@aag,{"***","aag"}],Style["aag",Bold],InputField[Dynamic@aag1,String,FieldSize-
>3],Checkbox[Dynamic@agg,{"***","agg"}],Style["agg",Bold],InputField[Dynamic@agg1,String,FieldSize->3]},
       {Checkbox[Dynamic@gtt,{"***","gtt"}],Style["gtt",Bold],InputField[Dynamic@gtt1,String,FieldSize-
>3],Checkbox[Dynamic@gct,{"***","gct"}],Style["gct",Bold],InputField[Dynamic@gct1,String,FieldSize-
>3],Checkbox[Dynamic@gat,{"***","gat"}],Style["gat",Bold],InputField[Dynamic@gat1,String,FieldSize-
>3],Checkbox[Dynamic@ggt,{"***","ggt"}],Style["ggt",Bold],InputField[Dynamic@ggt1,String,FieldSize->3]},
       {Checkbox[Dynamic@gtc,{"***","gtc"}],Style["gtc",Bold],InputField[Dynamic@gtc1,String,FieldSize-
>3],Checkbox[Dynamic@gcc,{"***","gcc"}],Style["gcc",Bold],InputField[Dynamic@gcc1,String,FieldSize-
>3],Checkbox[Dynamic@gac,{"***","gac"}],Style["gac",Bold],InputField[Dynamic@gac1,String,FieldSize-
>3],Checkbox[Dynamic@ggc,{"***","ggc"}],Style["ggc",Bold],InputField[Dynamic@ggc1,String,FieldSize->3]},
       {Checkbox[Dynamic@gta,{"***","gta"}],Style["gta",Bold],InputField[Dynamic@gta1,String,FieldSize-
>3],Checkbox[Dynamic@gca,{"***","gca"}],Style["gca",Bold],InputField[Dynamic@gca1,String,FieldSize-
>3],Checkbox[Dynamic@gaa,{"***","gaa"}],Style["gaa",Bold],InputField[Dynamic@gaa1,String,FieldSize-
>3],Checkbox[Dynamic@gga,{"***","gga"}],Style["gga",Bold],InputField[Dynamic@gga1,String,FieldSize->3]},
       {Checkbox[Dynamic@gtg,{"***","gtg"}],Style["gtg",Bold],InputField[Dynamic@gtg1,String,FieldSize-
>3],Checkbox[Dynamic@gcg,{"***","gcg"}],Style["gcg",Bold],InputField[Dynamic@gcg1,String,FieldSize-
>3],Checkbox[Dynamic@gag,{"***","gag"}],Style["gag",Bold],InputField[Dynamic@gag1,String,FieldSize-
>3],Checkbox[Dynamic@ggg,{"***","ggg"}],Style["ggg",Bold],InputField[Dynamic@ggg1,String,FieldSize->3]},
       { DefaultButton[DialogReturn[TranslationalTable={"ttt"->ToString@ttt1,"ttc"->ToString@ttc1,"tta"-
>ToString@tta1,"ttg"->ToString@ttg1,"tct"->ToString@tct1,"tcc"->ToString@tcc1,"tca"->ToString@tca1,"tcg"-
>ToString@tcg1,"taa"->ToString@taa1,"tat"->ToString@tat1,"tac"->ToString@tac1,"tag"->ToString@tag1,"tga"-
>ToString@tga1,"tgt"->ToString@tgt1,"tgc"->ToString@tgc1,"tgg"->ToString@tgg1,"ctt"->ToString@ctt1,"ctc"-
>ToString@ctc1,"cta"->ToString@cta1,"ctg"->ToString@ctg1,"cct"->ToString@cct1,"ccc"->ToString@ccc1,"cca"-
>ToString@cca1,"ccg"->ToString@ccg1,"cat"->ToString@cat1,"cac"->ToString@cac1,"caa"->ToString@caa1,"cag"-
>ToString@cag1,"cgt"->ToString@cgt1,"cgc"->ToString@cgc1,"cga"->ToString@cga1,"cgg"->ToString@cgg1,"att"-
>ToString@att1,"atc"->ToString@atc1,"ata"->ToString@ata1,"atg"->ToString@atg1,"act"->ToString@act1,"acc"-
>ToString@acc1,"aca"->ToString@aca1,"acg"->ToString@acg1,"aat"->ToString@aat1,"aac"->ToString@aac1,"aaa"-
>ToString@aaa1,"aag"->ToString@aag1,"agt"->ToString@agt1,"agc"->ToString@agc1,"aga"->ToString@aga1,"agg"-
>ToString@agg1,"gtt"->ToString@gtt1,"gtc"->ToString@gtc1,"gta"->ToString@gta1,"gtg"->ToString@gtg1,"gct"-
>ToString@gct1,"gcc"->ToString@gcc1,"gca"->ToString@gca1,"gcg"->ToString@gcg1,"gat"->ToString@gat1,"gac"-
>ToString@gac1,"gaa"->ToString@gaa1,"gag"->ToString@gag1,"ggt"->ToString@ggt1,"ggc"->ToString@ggc1,"gga"-
>ToString@gga1,"ggg"->ToString@ggg1};

RareCod=DeleteDuplicates@{ttt,ttc,tta,ttg,tct,tcc,tca,tcg,taa,tat,tac,tag,tga,tgt,tgc,tgg,ctt,ctc,cta,ctg,c
ct,ccc,cca,ccg,cat,cac,caa,cag,cgt,cgc,cga,cgg,att,atc,ata,atg,act,acc,aca,acg,aat,aac,aaa,aag,agt,agc,aga,
agg,gtt,gtc,gta,gtg,gct,gcc,gca,gcg,gat,gac,gaa,gag,ggt,ggc,gga,ggg}]],SpanFromLeft,SpanFromLeft,SpanFromLe
ft,SpanFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft,SpanFromLeft}
       },
      Dividers->{{1->True,4->True,7->True,10->True,13->True},All}]},
   WindowTitle->"Translational Table"]],
 TranslationalTable;RareCod;]
```

Interface for Amino Acids Selection

```
Interpretation[{TotalAminoAcid={"Ala","Arg","Asn","Asp","Cys","Gln","Glu","Gly","His","Ile","Leu","Lys","Me
t","Phe","Pro","Ser","Thr","Trp","Tyr","Val"}},
 Panel@Grid[{
    {Style["Which amino acid(s) do you want to include in your primers?",Bold,Medium],SpanFromLeft},
        {TogglerBar[Dynamic[AminoAcid],
{"Ala","Arg","Asn","Asp","Cys","Gln","Glu","Gly","His","Ile","Leu","Lys","Met","Phe","Pro","Ser","Thr","Trp
","Tyr","Val"}],SpanFromLeft},{ Button["Select
All",AminoAcid={"Ala","Arg","Asn","Asp","Cys","Gln","Glu","Gly","His","Ile","Leu","Lys","Met","Phe","Pro","
Ser","Thr","Trp","Tyr","Val"}],
     Button["Remove All",AminoAcid={}]},
    {Style["How many primers do you want to mix?",Bold,Medium],CheckboxBar[Dynamic[MixNo],{2,3,4,5}]}
    },Dividers->{{1,None,1},{1,None,None,1,1,1}},ItemSize-
>All],AminoAcid//Sort;ExcludedAminoAcid=Complement[TotalAminoAcid,AminoAcid]//Sort;MixNo;Panel[Grid[{{Style
["Included Amino Acid(s):", Bold],Row[AminoAcid,","]},{Style["Excluded Amino Acid(s):",
Bold],Row[ExcludedAminoAcid,","]}},Alignment->Left]]]
```

Chapter V – Appendix

Core Software

The code below is dependent on the two graphic user interfaces generated by the two codes in the sections above.

```
 Nuc={{"A",{"a"}},{"C",{"c"}},{"G",{"g"}},{"T",{"t"}},{"R",{"a","g"}},{"Y",{"c","t"}},{"S",{"c","g"}},{"W",
{"a","t"}},{"K",{"g","t"}},{"M",{"a","c"}},{"B",{"c","g","t"}},{"D",{"a","g","t"}},{"H",{"a","c","t"}},
{"V",{"a","c","g"}},{"N",{"a","c","g","t"}}};

Transl=Dispatch[TranslationalTable];

PrimerTot=Flatten[Table[{{Nuc[[i,1]]<>Nuc[[j,1]]<>Nuc[[k,1]]},Flatten[Outer[StringJoin,
#,Nuc[[k,2]]]&/@Outer[StringJoin, Nuc[[i,2]], Nuc[[j,2]]],2]},{i,1,15},{j,1,15},{k,1,15}],2];(*Generates
all the possible primers using the IUPAC alphabet*)

Nprimertot=Dimensions[PrimerTot];
PrimerNoRareCodon=Table[If[Intersection[PrimerTot[[i,2]],RareCod]==={},PrimerTot[[i]],Nothing],
{i,1,Nprimertot[[1]]}];(*Removes from the list of primers the ones that code for rare codons*)

Nprimernorarecodon=Dimensions[PrimerNoRareCodon];

Table[{PrimerNoRareCodon[[i,1]],PrimerNoRareCodon[[i,2]],PrimerNoRareCodon[[i,2]]/.Transl},
{i,1,Nprimernorarecodon[[1]]}];(*Adds to the previuous table the list of the amino acids they encode for*)

PrimerNoRedundant=Table[If[CountDistinct[Part[#,2]&/@Tally[%[[i,3]]]]==1,%[[i]],Nothing],{i,1,Dimensions[%]
[[1]]}];(*keeps only the primers that cover evenly a given set of amino acid, that means each one either
once or twice etc.*)

m=Length[AminoAcid];
FinalPrimer=Table[If[Intersection[PrimerNoRedundant[[i,3]],ExcludedAminoAcid]==={},
{PrimerNoRedundant[[i,1]],PrimerNoRedundant[[i,2]],DeleteDuplicates@PrimerNoRedundant[[i,3]],
{CountDistinct@DeleteDuplicates@PrimerNoRedundant[[i,3]]}},Nothing],{i,1,Dimensions[PrimerNoRedundant]
[[1]]}];(*Generate the final list of non redundant amino acids that do not contain rare codon AND do not
code for the amino acids unselected by the user*)

Nfinalprimer=Dimensions[FinalPrimer];

PrimerL=Sort[DeleteDuplicates[Table[CountDistinct[FinalPrimer[[i,3]]],{i,1,Nfinalprimer[[1]]}]],Smaller];
Npl=Dimensions[PrimerL][[1]];
a[x_]:=Table[If[CountDistinct[FinalPrimer[[j,3]]]==x,FinalPrimer[[j]],Nothing],{j,1,Nfinalprimer[[1]]}];
For[x=1,x<=Max[PrimerL],x++,a[x]];(*Divides the list of primers based on their degeneracy*)

g[i_]:=Gather[a[i],(Sort@#1[[3]])==Sort@(#2[[3]])&];(*pools the primers of the same length that code for
the same set of amino acids*)

For[i=1,i<=Max[PrimerL],i++,Eq[i]=Map[First,g[i],{2}];Signpost[i]=First/@g[i];
 aa[i]=Part[Dimensions[Signpost[i]],1]];

blk[i_]:=blk[i]=If[CountDistinct@#[[3]]==i[[1]]*i[[2]],#,Nothing]&/@Apply[Join,Map[Transpose,Subsets[Signpo
st[i[[1]]],{i[[2]]}]],{2}];

combine[x_,y_]:=combine[x,y]=If[CountDistinct[#[[3]]]===(CountDistinct@x[[1,3]]
+CountDistinct@y[[1,3]]),#,Nothing]&/@Apply[Join,Transpose/@Flatten[Outer[List,x,y,1],1],{2}]

z2=Flatten[Table[If[(PrimerL[[i]]+PrimerL[[j]])===m,{PrimerL[[i]],PrimerL[[j]]},Nothing],{i,1,Npl},
{j,i,Npl}],1];
z3=Flatten[Table[If[(PrimerL[[i]]+PrimerL[[j]]+PrimerL[[k]])===m,
{PrimerL[[i]],PrimerL[[j]],PrimerL[[k]]},Nothing],{i,1,Npl},{j,i,Npl},{k,j,Npl}],2];
z4=Flatten[Table[If[(PrimerL[[i]]+PrimerL[[j]]+PrimerL[[k]]+PrimerL[[l]])===m,
{PrimerL[[i]],PrimerL[[j]],PrimerL[[k]],PrimerL[[l]]},Nothing],{i,1,Npl},{j,i,Npl},{k,j,Npl},{l,k,Npl}],3];
z5=Flatten[Table[If[(PrimerL[[i]]+PrimerL[[j]]+PrimerL[[k]]+PrimerL[[l]]+PrimerL[[o]])===m,
{PrimerL[[i]],PrimerL[[j]],PrimerL[[k]],PrimerL[[l]],PrimerL[[o]]},Nothing],{i,1,Npl},{j,i,Npl},{k,j,Npl},
{l,k,Npl},{o,l,Npl}],4];

w2=Reverse/@(Tally/@z2);
w3=Reverse/@(Tally/@z3);
w4=Reverse/@(Tally/@z4);
w5=Reverse/@(Tally/@z5);

out[2]=If[MemberQ[MixNo,2],Flatten[ParallelMap[Fold[combine,#]&,Map[blk,w2,{2}]],1],"Mixes of 2 primers
where not searched"];
out[3]=If[MemberQ[MixNo,3],Flatten[ParallelMap[Fold[combine,#]&,Map[blk,w3,{2}]],1],"Mixes of 3 primers
where not searched"];
out[4]=If[MemberQ[MixNo,4],Flatten[ParallelMap[Fold[combine,#]&,Map[blk,w4,{2}]],1],"Mixes of 4 primers
where not searched"];
out[5]=If[MemberQ[MixNo,5],Flatten[ParallelMap[Fold[combine,#]&,Map[blk,w5,{2}]],1],"Mixes of 5 primers
where not searched"];

Panel@Grid[{{Style["Mix of 2 primers",Bold],If[MemberQ[MixNo,2],If[out[2]==={},"No solutions
```

```
found",TableForm[Join@@@(j[2]=Transpose@Transpose[out[2]][[{1,4}]]),TableSpacing->{1,1}]],out[2]]},
   {Style["Mix of 3 primers",Bold],If[MemberQ[MixNo,3],If[out[3]==={},"No solutions
found",TableForm[Join@@@(j[3]=Transpose@Transpose[out[3]][[{1,4}]]),TableSpacing->{1,1}]],out[3]]},
   {Style["Mix of 4 primers",Bold],If[MemberQ[MixNo,4],If[out[4]==={},"No solutions
found",TableForm[Join@@@(j[4]=Transpose@Transpose[out[4]][[{1,4}]]),TableSpacing->{1,1}]],out[4]]},
   {Style["Mix of 5 primers",Bold],If[MemberQ[MixNo,5],If[out[5]==={},"No solutions
found",TableForm[Join@@@(j[5]=Transpose@Transpose[out[5]][[{1,4}]]),TableSpacing->{1,1}]],out[5]]}}]

TableofEquivalence[x_]:=(f[x]=DeleteDuplicates@Flatten[Transpose/@j[x],1];t[x]=Table[Position[Eq[f[x]
[[i,2]]],f[x][[i,1]]][[1,1]],{i,Dimensions[f[x]][[1]]}];SortBy[Table[Eq[f[x][[i,2]]][[t[x][[i]]]],
{i,Dimensions[t[x]][[1]]}],#1[[1]]&])/;MemberQ[MixNo,x]&&out[x]=!={};

If[out[#]=!={},#,Nothing]&/@MixNo;
Panel@TableForm[DeleteCases[Flatten[Union@(TableofEquivalence/@%),1],_?(Length[#]===1&)],TableSpacing-
>{.5,1}]

Clear["Global`*"]
```

Acknowledgements

At last, my journey as a Ph.D. student has come to an end. These years have been more valuable for my personal development than I can acknowledge now. They have been years full of joy, but also years during which I had to come to terms with something university doesn't train you for: failure. I realised that the impression of scientists that books create in your head can be very deceiving. What books never teach you is that the life of a scientist is a constant fight against mistakes, misinterpretations, unexplainable phenomena and mysterious events which mar the beauty of your hypotheses about how things should be, but aren't. And I must admit that, after all, I probably had an easy time. Nonetheless, in several occasions I struggled to see the light at the end of the tunnel.

I would like to thank all the people that gave me the chance embark on this journey, and all the great people I met along the way, for everything they did for me, for their presence, their support, their smiles, for giving me the strength I needed to complete this journey, that would have felt so much scarier otherwise.

I would like to thank Jason, my supervisor, for giving me the opportunity to complete a successful Ph.D. in this very prestigious institute. I would also like to thank Emma, for being a fantastic admin and, more importantly, a good friend.

Along my journey I also learn how good things often come to an end. I now know that all these

lessons, for how much unpleasant they might seem, need to be learnt sooner or later, because they are part of life. I wish to sincerely thank all the people that were by my side in the darkest days, in particular Edoardo, Lisa, Jakob, Charlie, Florence, Wesley and Anne, Julian, Jing, my parents Eleonora and Francesco, my aunt Mariella and my sister Sonia.

In spite of circumstances, I wish to thank Michele, for what he meant to me for a very big part of my life, and everybody who made me who I am today. And lastly, I wish to thank Bruno, for making me feel whole again after a long time and for reminding me that sometimes pearls can be found where you would least expect them.

Bibliography

1. Avery, O.T., Macleod, C.M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79**, 137-158 (1944).
2. Hershey, A.D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**, 39-56 (1952).
3. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
4. Crick, F.H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138-163 (1958).
5. Nirenberg, M.W. & Matthaei, J.H. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A* **47**, 1588-1602 (1961).
6. Jones, O.W., Jr. & Nirenberg, M.W. Degeneracy in the amino acid code. *Biochim Biophys Acta* **119**, 400-406 (1966).
7. Hoagland, M.B., Keller, E.B. & Zamecnik, P.C. Enzymatic carboxyl activation of amino acids. *J Biol Chem* **218**, 345-358 (1956).
8. Hoagland, M.B., Stephenson, M.L., Scott, J.F., Hecht, L.I. & Zamecnik, P.C. A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem* **231**, 241-257 (1958).
9. Zamecnik, P. From protein synthesis to genetic insertion. *Annu Rev Biochem* **74**, 1-28 (2005).
10. Weiss, S.B. & Gladstone, L. A Mammalian System for the Incorporation of Cytidine Triphosphate into Ribonucleic Acid. *Journal of the American Chemical Society* **81**, 4118-4119

(1959).

11. Landick, R. A long time in the making--the Nobel Prize for RNA polymerase. *Cell* **127**, 1087-1090 (2006).
12. Ramakrishnan, V. Ribosome structure and the mechanism of translation. *Cell* **108**, 557-572 (2002).
13. Lyons, S.M., Fay, M.M. & Ivanov, P. The role of RNA modifications in the regulation of tRNA cleavage. *FEBS Lett* **592**, 2828-2844 (2018).
14. Chan, P.P. & Lowe, T.M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* **37**, D93-97 (2009).
15. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. & Steinberg, S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res* **26**, 148-153 (1998).
16. Savage, D.F., de Crecy-Lagard, V. & Bishop, A.C. Molecular determinants of dihydrouridine synthase activity. *FEBS Lett* **580**, 5198-5202 (2006).
17. Shi, H. & Moore, P.B. The crystal structure of yeast phenylalanine tRNA at 1.93 A resolution: a classic structure revisited. *RNA* **6**, 1091-1105 (2000).
18. Giege, R. et al. Structure of transfer RNAs: similarity and variability. *Wiley Interdiscip Rev RNA* **3**, 37-61 (2012).
19. Juhling, T. et al. Small but large enough: structural properties of armless mitochondrial tRNAs from the nematode Romanomermis culicivorax. *Nucleic Acids Res* **46**, 9170-9180 (2018).
20. Giegé, R., Sissler, M. & Florentz, C. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Research* **26**, 5017-5035 (1998).
21. Ibba, M. & Soll, D. Aminoacyl-tRNA synthesis. *Annual Review of Biochemistry* **69**, 617-650 (2000).
22. Peacock, J.R. et al. Amino acid-dependent stability of the acyl linkage in aminoacyl-tRNA. *RNA* **20**, 758-764 (2014).
23. Zaher, H.S. & Green, R. Fidelity at the molecular level: lessons from protein synthesis. *Cell* **136**, 746-762 (2009).
24. Kaiser, F. et al. Characterization of Amino Acid Recognition in Aminoacyl-tRNA Synthetases. *bioRxiv*, 606459 (2019).
25. Fukai, S. et al. Structural basis for double-sieve discrimination of L-valine from L-isoleucine and L-threonine by the complex of tRNA(Val) and valyl-tRNA synthetase. *Cell* **103**, 793-803 (2000).
26. Abe, T. et al. tRNADB-CE 2011: tRNA gene database curated manually by experts. *Nucleic Acids Res* **39**, D210-213 (2011).
27. Osawa, S., Jukes, T.H., Watanabe, K. & Muto, A. Recent evidence for evolution of the genetic code. *Microbiol Rev* **56**, 229-264 (1992).
28. Chin, J.W. Expanding and reprogramming the genetic code. *Nature* **550**, 53-60 (2017).
29. Ieong, K.W., Pavlov, M.Y., Kwiatkowski, M., Forster, A.C. & Ehrenberg, M. Inefficient delivery but fast peptide bond formation of unnatural L-aminoacyl-tRNAs in translation. *J Am Chem Soc* **134**, 17955-17962 (2012).
30. Fredens, J. et al. Total synthesis of Escherichia coli with a recoded genome. *Nature* **569**, 514-518 (2019).
31. Young, D.D. & Schultz, P.G. Playing with the Molecules of Life. *ACS Chem Biol* **13**, 854-870

(2018).

32. Uttamapinant, C. et al. Genetic code expansion enables live-cell and super-resolution imaging of site-specifically labeled cellular proteins. *J Am Chem Soc* **137**, 4602-4605 (2015).

33. Elliott, T.S., Bianco, A., Townsley, F.M., Fried, S.D. & Chin, J.W. Tagging and Enriching Proteins Enables Cell-Specific Proteomics. *Cell Chem Biol* **23**, 805-815 (2016).

34. Wilkins, B.J. et al. A cascade of histone modifications induces chromatin condensation in mitosis. *Science* **343**, 77-80 (2014).

35. Yang, Y. et al. Genetically encoded protein photocrosslinker with a transferable mass spectrometry-identifiable label. *Nat Commun* **7**, 12299 (2016).

36. Yang, T., Li, X.M., Bao, X., Fung, Y.M. & Li, X.D. Photo-lysine captures proteins that bind lysine post-translational modifications. *Nat Chem Biol* **12**, 70-72 (2016).

37. Hemphill, J., Chou, C., Chin, J.W. & Deiters, A. Genetically encoded light-activated transcription for spatiotemporal control of gene expression and gene silencing in mammalian cells. *J Am Chem Soc* **135**, 13433-13439 (2013).

38. Klippenstein, V., Hoppmann, C., Ye, S., Wang, L. & Paoletti, P. Optocontrol of glutamate receptor activity by single side-chain photoisomerization. *Elife* **6** (2017).

39. Ubersax, J.A. & Ferrell, J.E., Jr. Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol* **8**, 530-541 (2007).

40. Park, H.S. et al. Expanding the genetic code of Escherichia coli with phosphoserine. *Science* **333**, 1151-1154 (2011).

41. Rogerson, D.T. et al. Efficient genetic encoding of phosphoserine and its nonhydrolyzable analog. *Nat Chem Biol* **11**, 496-503 (2015).

42. Zhang, M.S. et al. Biosynthesis and genetic encoding of phosphothreonine through parallel selection and deep sequencing. *Nat Methods* **14**, 729-736 (2017).

43. Umehara, T. et al. N-acetyl lysyl-tRNA synthetases evolved by a CcdB-based selection possess N-acetyl lysine specificity in vitro and in vivo. *FEBS Lett* **586**, 729-733 (2012).

44. Nguyen, D.P., Garcia Alai, M.M., Virdee, S. & Chin, J.W. Genetically directing varepsilon-N, N-dimethyl-L-lysine in recombinant histones. *Chem Biol* **17**, 1072-1076 (2010).

45. Stahl, F.W. The amber mutants of phage T4. *Genetics* **141**, 439-442 (1995).

46. Garen, A., Garen, S. & Wilhelm, R.C. Suppressor genes for nonsense mutations. I. The Su-1, Su-2 and Su-3 genes of Escherichia coli. *J Mol Biol* **14**, 167-178 (1965).

47. Eggertsson, G. & Soll, D. Transfer ribonucleic acid-mediated suppression of termination codons in Escherichia coli. *Microbiol Rev* **52**, 354-374 (1988).

48. Nakamura, Y., Gojobori, T. & Ikemura, T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* **28**, 292 (2000).

49. Furter, R. Expansion of the genetic code: site-directed p-fluoro-phenylalanine incorporation in Escherichia coli. *Protein Sci* **7**, 419-426 (1998).

50. Liu, D.R., Magliery, T.J., Pastrnak, M. & Schultz, P.G. Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins in vivo. *Proc Natl Acad Sci U S A* **94**, 10092-10097 (1997).

51. Bult, C.J. et al. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science* **273**, 1058-1073 (1996).

52. Wang, L., Magliery, T.J., Liu, D.R. & Schultz, P.G. A new functional suppressor

tRNA/aminoacyl-tRNA synthetase pair for the in vivo incorporation of unnatural amino acids into proteins. *Journal of the American Chemical Society* **122**, 5010-5011 (2000).

53. Wang, L., Brock, A., Herberich, B. & Schultz, P.G. Expanding the genetic code of Escherichia coli. *Science* **292**, 498-500 (2001).

54. Liu, C.C. & Schultz, P.G. Adding new chemistries to the genetic code. *Annu Rev Biochem* **79**, 413-444 (2010).

55. Suzuki, T. et al. Crystal structures reveal an elusive functional domain of pyrrolysyl-tRNA synthetase. *Nat Chem Biol* **13**, 1261-1266 (2017).

56. Burke, S.A., Lo, S.L. & Krzycki, J.A. Clustered genes encoding the methyltransferases of methanogenesis from monomethylamine. *J Bacteriol* **180**, 3432-3440 (1998).

57. Wan, W., Tharp, J.M. & Liu, W.R. Pyrrolysyl-tRNA synthetase: an ordinary enzyme but an outstanding genetic code expansion tool. *Biochim Biophys Acta* **1844**, 1059-1070 (2014).

58. Hao, B. et al. A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science* **296**, 1462-1466 (2002).

59. Krzycki, J.A. The path of lysine to pyrrolysine. *Curr Opin Chem Biol* **17**, 619-625 (2013).

60. Srinivasan, G., James, C.M. & Krzycki, J.A. Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science* **296**, 1459-1462 (2002).

61. Nozawa, K. et al. Pyrrolysyl-tRNA synthetase-tRNA(Pyl) structure reveals the molecular basis of orthogonality. *Nature* **457**, 1163-1167 (2009).

62. Polycarpo, C.R. et al. Pyrrolysine analogues as substrates for pyrrolysyl-tRNA synthetase. *FEBS Lett* **580**, 6695-6700 (2006).

63. Borrel, G. et al. Unique characteristics of the pyrrolysine system in the 7th order of methanogens: implications for the evolution of a genetic code expansion cassette. *Archaea* **2014**, 374146 (2014).

64. Willis, J.C.W. & Chin, J.W. Mutually orthogonal pyrrolysyl-tRNA synthetase/tRNA pairs. *Nat Chem* **10**, 831-837 (2018).

65. Ambrogelly, A. et al. Cys-tRNACys formation and cysteine biosynthesis in methanogenic archaea: two faces of the same problem? *Cell Mol Life Sci* **61**, 2437-2445 (2004).

66. Sauerwald, A. et al. RNA-dependent cysteine biosynthesis in archaea. *Science* **307**, 1969-1972 (2005).

67. Kamtekar, S. et al. Toward understanding phosphoseryl-tRNACys formation: the crystal structure of Methanococcus maripaludis phosphoseryl-tRNA synthetase. *Proc Natl Acad Sci U S A* **104**, 2620-2625 (2007).

68. Bennett, B.D. et al. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. *Nat Chem Biol* **5**, 593-599 (2009).

69. Beranek, V. et al. Genetically Encoded Protein Phosphorylation in Mammalian Cells. *Cell Chem Biol* **25**, 1067-1074 e1065 (2018).

70. Johnson, D.B. et al. RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. *Nat Chem Biol* **7**, 779-786 (2011).

71. Odoi, K.A., Huang, Y., Rezenom, Y.H. & Liu, W.R. Nonsense and sense suppression abilities of original and derivative Methanosarcina mazei pyrrolysyl-tRNA synthetase-tRNA(Pyl) pairs in the Escherichia coli BL21(DE3) cell strain. *PLoS One* **8**, e57035 (2013).

72. Curran, J.F. & Yarus, M. Reading frame selection and transfer RNA anticodon loop stacking. *Science* **238**, 1545-1550 (1987).

Bibliography

73. Hohsaka, T., Kajihara, D., Ashizuka, Y., Murakami, H. & Sisido, M. Efficient Incorporation of Nonnatural Amino Acids with Large Aromatic Groups into Streptavidin in In Vitro Protein Synthesizing Systems. *Journal of the American Chemical Society* **121**, 34-40 (1999).

74. Hohsaka, T., Ashizuka, Y., Taira, H., Murakami, H. & Sisido, M. Incorporation of nonnatural amino acids into proteins by using various four-base codons in an Escherichia coli in vitro translation system. *Biochemistry* **40**, 11060-11064 (2001).

75. Kajihara, D. et al. FRET analysis of protein conformational change through position-specific incorporation of fluorescent amino acids. *Nat Methods* **3**, 923-929 (2006).

76. Anderson, J.C. et al. An expanded genetic code with a functional quadruplet codon. *Proc Natl Acad Sci U S A* **101**, 7566-7571 (2004).

77. Rackham, O. & Chin, J.W. A network of orthogonal ribosome x mRNA pairs. *Nat Chem Biol* **1**, 159-166 (2005).

78. Wang, K., Neumann, H., Peak-Chew, S.Y. & Chin, J.W. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nat Biotechnol* **25**, 770-777 (2007).

79. Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. & Chin, J.W. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **464**, 441-444 (2010).

80. Switzer, C., Moroney, S.E. & Benner, S.A. Enzymatic Incorporation of a New Base Pair into DNA and Rna. *Journal of the American Chemical Society* **111**, 8322-8323 (1989).

81. Bain, J.D., Switzer, C., Chamberlin, A.R. & Benner, S.A. Ribosome-mediated incorporation of a non-standard amino acid into a peptide through expansion of the genetic code. *Nature* **356**, 537-539 (1992).

82. Yang, Z., Chen, F., Alvarado, J.B. & Benner, S.A. Amplification, mutation, and sequencing of a six-letter synthetic genetic system. *J Am Chem Soc* **133**, 15105-15112 (2011).

83. Hirao, I. et al. An unnatural hydrophobic base pair system: site-specific incorporation of nucleotide analogs into DNA and RNA. *Nat Methods* **3**, 729-735 (2006).

84. Kimoto, M., Kawai, R., Mitsui, T., Yokoyama, S. & Hirao, I. An unnatural base pair system for efficient PCR amplification and functionalization of DNA molecules. *Nucleic Acids Res* **37**, e14 (2009).

85. Malyshev, D.A. et al. Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. *Proc Natl Acad Sci U S A* **109**, 12005-12010 (2012).

86. Zhang, Y. et al. A semi-synthetic organism that stores and retrieves increased genetic information. *Nature* **551**, 644-647 (2017).

87. Mukai, T. et al. Codon reassignment in the Escherichia coli genetic code. *Nucleic Acids Res* **38**, 8188-8195 (2010).

88. Craigen, W.J., Cook, R.G., Tate, W.P. & Caskey, C.T. Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. *Proc Natl Acad Sci U S A* **82**, 3616-3620 (1985).

89. Wang, H.H. et al. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894-898 (2009).

90. Isaacs, F.J. et al. Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* **333**, 348-353 (2011).

91. Lajoie, M.J. et al. Genomically recoded organisms expand biological functions. *Science* **342**, 357-360 (2013).

92. Wang, K. et al. Defining synonymous codon compression schemes by genome recoding. *Nature* **539**, 59-64 (2016).

93. Pastrnak, M., Magliery, T.J. & Schultz, P.G. A New Orthogonal Suppressor tRNA/Aminoacyl-tRNA Synthetase Pair for Evolving an Organism with an Expanded Genetic Code. *Helvetica Chimica Acta* **83**, 2277-2286 (2000).

94. Anderson, J.C. & Schultz, P.G. Adaptation of an orthogonal archaeal leucyl-tRNA and synthetase pair for four-base, amber, and opal suppression. *Biochemistry* **42**, 9598-9608 (2003).

95. Zambaldo, C. et al. An orthogonal seryl-tRNA synthetase/tRNA pair for noncanonical amino acid mutagenesis in Escherichia coli. *Bioorg Med Chem* **28**, 115662 (2020).

96. Chatterjee, A., Xiao, H. & Schultz, P.G. Evolution of multiple, mutually orthogonal prolyl-tRNA synthetase/tRNA pairs for unnatural amino acid mutagenesis in Escherichia coli. *Proc Natl Acad Sci U S A* **109**, 14841-14846 (2012).

97. Santoro, S.W., Anderson, J.C., Lakshman, V. & Schultz, P.G. An archaebacteria-derived glutamyl-tRNA synthetase and tRNA pair for unnatural amino acid mutagenesis of proteins in Escherichia coli. *Nucleic Acids Res* **31**, 6700-6709 (2003).

98. Italia, J.S. et al. An orthogonalized platform for genetic code expansion in both bacteria and eukaryotes. *Nat Chem Biol* **13**, 446-450 (2017).

99. Hughes, R.A. & Ellington, A.D. Rational design of an orthogonal tryptophanyl nonsense suppressor tRNA. *Nucleic Acids Res* **38**, 6813-6830 (2010).

100. Italia, J.S. et al. Mutually Orthogonal Nonsense-Suppression Systems and Conjugation Chemistries for Precise Protein Labeling at up to Three Distinct Sites. *J Am Chem Soc* **141**, 6204-6212 (2019).

101. Wan, W. et al. A facile system for genetic incorporation of two different noncanonical amino acids into one protein in Escherichia coli. *Angew Chem Int Ed Engl* **49**, 3211-3214 (2010).

102. Chatterjee, A., Sun, S.B., Furman, J.L., Xiao, H. & Schultz, P.G. A versatile platform for single- and multiple-unnatural amino acid mutagenesis in Escherichia coli. *Biochemistry* **52**, 1828-1837 (2013).

103. Wang, K. et al. Optimized orthogonal translation of unnatural amino acids enables spontaneous protein double-labelling and FRET. *Nat Chem* **6**, 393-403 (2014).

104. Sachdeva, A., Wang, K., Elliott, T. & Chin, J.W. Concerted, rapid, quantitative, and site-specific dual labeling of proteins. *J Am Chem Soc* **136**, 7785-7788 (2014).

105. Zheng, Y., Gilgenast, M.J., Hauc, S. & Chatterjee, A. Capturing Post-Translational Modification-Triggered Protein-Protein Interactions Using Dual Noncanonical Amino Acid Mutagenesis. *ACS Chem Biol* **13**, 1137-1141 (2018).

106. Yuan, J., Gogakos, T., Babina, A.M., Soll, D. & Randau, L. Change of tRNA identity leads to a divergent orthogonal histidyl-tRNA synthetase/tRNAHis pair. *Nucleic Acids Res* **39**, 2286-2293 (2011).

107. Varshney, U., Lee, C.P. & RajBhandary, U.L. Direct analysis of aminoacylation levels of tRNAs in vivo. Application to studying recognition of Escherichia coli initiator tRNA mutants by glutaminyl-tRNA synthetase. *J Biol Chem* **266**, 24712-24718 (1991).

108. Borner, G.V., Morl, M., Janke, A. & Paabo, S. RNA editing changes the identity of a mitochondrial tRNA in marsupials. *EMBO J* **15**, 5949-5957 (1996).

109. Gaston, K.W., Rubio, M.A. & Alfonzo, J.D. OXOPAP assay: for selective amplification of aminoacylated tRNAs from total cellular fractions. *Methods* **44**, 170-175 (2008).

110. Hunt, J.A. Terminal-Sequence Studies of High-Molecular-Weight Ribonucleic. The Reaction of

Periodate-Oxidized Ribonucleosides , 5'-Ribonucleotides and Ribonucleic Acid with Isoniazid. *Biochem J* **95**, 541-551 (1965).

111. Wittig, B. & Wittig, S. Reverse transcription of tRNA. *Nucleic Acids Res* **5**, 1165-1178 (1978).

112. Steer, B.A. & Schimmel, P. Major Anticodon-binding Region Missing from an Archaebacterial tRNA Synthetase. *Journal of Biological Chemistry* **274**, 35601-35606 (1999).

113. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).

114. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**, 11-16 (2004).

115. Kinouchi, M. & Kurokawa, K. tRNAfinder: A Software System To Find All tRNA Genes in the DNA Sequence Based on the Cloverleaf Secondary Structure. *Journal of Computer Aided Chemistry* **7**, 116-124 (2006).

116. Hou, Y.M. CCA addition to tRNA: implications for tRNA quality control. *IUBMB Life* **62**, 251-260 (2010).

117. Lee, S.W., Cho, B.H., Park, S.G. & Kim, S. Aminoacyl-tRNA synthetase complexes: beyond translation. *J Cell Sci* **117**, 3725-3734 (2004).

118. Quinn, C.L., Tao, N. & Schimmel, P. Species-specific microhelix aminoacylation by a eukaryotic pathogen tRNA synthetase dependent on a single base pair. *Biochemistry* **34**, 12489-12495 (1995).

119. Larkin, D.C., Williams, A.M., Martinis, S.A. & Fox, G.E. Identification of essential domains for Escherichia coli tRNA(leu) aminoacylation and amino acid editing using minimalist RNA molecules. *Nucleic Acids Res* **30**, 2103-2113 (2002).

120. Aldinger, C.A., Leisinger, A.K. & Igloi, G.L. The influence of identity elements on the aminoacylation of tRNA(Arg) by plant and Escherichia coli arginyl-tRNA synthetases. *FEBS J* **279**, 3622-3638 (2012).

121. Pedelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C. & Waldo, G.S. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79-88 (2006).

122. Inokuchi, H., Yamao, F., Sakano, H. & Ozeki, H. Identification of transfer RNA suppressors in Escherichia coli. I. Amber suppressor su+2, an anticodon mutant of tRNA2Gln. *J Mol Biol* **132**, 649-662 (1979).

123. Briand, C. et al. An intermediate step in the recognition of tRNA(Asp) by aspartyl-tRNA synthetase. *J Mol Biol* **299**, 1051-1060 (2000).

124. Stemmer, W.P. & Morris, S.K. Enzymatic inverse PCR: a restriction site independent, single-fragment method for high-efficiency, site-directed mutagenesis. *Biotechniques* **13**, 214-220 (1992).

125. Packer, M.J., Dauncey, M.P. & Hunter, C.A. Sequence-dependent DNA structure: tetranucleotide conformational maps. *J Mol Biol* **295**, 85-103 (2000).

126. Eiler, S., Dock-Bregeon, A., Moulinier, L., Thierry, J.C. & Moras, D. Synthesis of aspartyl-tRNA(Asp) in Escherichia coli--a snapshot of the second step. *EMBO J* **18**, 6532-6541 (1999).

127. Cavarelli, J. et al. The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction. *EMBO J* **13**, 327-337 (1994).

128. Tang, L. et al. Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. *Biotechniques* **52**, 149-158 (2012).

129. Hauenstein, S., Zhang, C.M., Hou, Y.M. & Perona, J.J. Shape-selective RNA recognition by

cysteinyl-tRNA synthetase. *Nat Struct Mol Biol* **11**, 1134-1141 (2004).

130. Newberry, K.J., Hou, Y.M. & Perona, J.J. Structural origins of amino acid selection without editing by cysteinyl-tRNA synthetase. *EMBO J* **21**, 2778-2787 (2002).

131. Arnez, J.G. & Steitz, T.A. Crystal structures of three misacylating mutants of Escherichia coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP. *Biochemistry* **35**, 14725-14733 (1996).

132. Liu, D.R. & Schultz, P.G. Progress toward the evolution of an organism with an expanded genetic code. *Proc Natl Acad Sci U S A* **96**, 4780-4785 (1999).

133. Ito, T. & Yokoyama, S. Two enzymes bound to one transfer RNA assume alternative conformations for consecutive reactions. *Nature* **467**, 612-616 (2010).

134. Qin, X. et al. Cocrystal structures of glycyl-tRNA synthetase in complex with tRNA suggest multiple conformational states in glycylation. *J Biol Chem* **289**, 20359-20369 (2014).

135. Tian, Q., Wang, C., Liu, Y. & Xie, W. Structural basis for recognition of G-1-containing tRNA by histidyl-tRNA synthetase. *Nucleic Acids Res* **43**, 2980-2990 (2015).

136. Fukunaga, J., Yokogawa, T., Ohno, S. & Nishikawa, K. Misacylation of yeast amber suppressor tRNA(Tyr) by E. coli lysyl-tRNA synthetase and its effective repression by genetic engineering of the tRNA sequence. *J Biochem* **139**, 689-696 (2006).

137. Kobayashi, T. et al. Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. *Nat Struct Biol* **10**, 425-432 (2003).

138. Kuratani, M. et al. Crystal structures of tyrosyl-tRNA synthetases from Archaea. *J Mol Biol* **355**, 395-408 (2006).

139. Chin, J.W., Martin, A.B., King, D.S., Wang, L. & Schultz, P.G. Addition of a photocrosslinking amino acid to the genetic code of Escherichia coli. *Proc Natl Acad Sci U S A* **99**, 11020-11024 (2002).

140. Chin, J.W. et al. Addition of p-azido-L-phenylalanine to the genetic code of Escherichia coli. *J Am Chem Soc* **124**, 9026-9027 (2002).

141. Brustad, E. et al. A genetically encoded boronate-containing amino acid. *Angew Chem Int Ed Engl* **47**, 8220-8223 (2008).

142. Xie, J. et al. The site-specific incorporation of p-iodo-L-phenylalanine into proteins for structure determination. *Nat Biotechnol* **22**, 1297-1301 (2004).

143. Cooley, R.B. et al. Structural basis of improved second-generation 3-nitro-tyrosine tRNA synthetases. *Biochemistry* **53**, 1916-1924 (2014).

144. Remington, S.J. Green fluorescent protein: a perspective. *Protein Sci* **20**, 1509-1519 (2011).

145. Heim, R., Prasher, D.C. & Tsien, R.Y. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proc Natl Acad Sci U S A* **91**, 12501-12504 (1994).

146. Strack, R.L. et al. A rapidly maturing far-red derivative of DsRed-Express2 for whole-cell labeling. *Biochemistry* **48**, 8279-8281 (2009).

147. Reddington, S.C. et al. Different photochemical events of a genetically encoded phenyl azide define and modulate GFP fluorescence. *Angew Chem Int Ed Engl* **52**, 5974-5977 (2013).

148. Ostrov, N. et al. Design, synthesis, and testing toward a 57-codon genome. *Science* **353**, 819-822 (2016).

149. Brubaker, L.H. & McCorquodale, D.J. The Preparation of Amino Acid-Transfer Ribonucleic Acid from Escherichia Coli by Direct Phenol Extraction of Intact Cells. *Biochim Biophys Acta* **76**, 48-53 (1963).

Table of Contents

Figures

Tables